Univerza v Ljubljani

Fakulteta za elektrotehniko

Nikola Marić

UPORABA VEČMODALNIH VELIKIH JEZIKOVNIH MODELOV ZA ZAZNAVANJE NAPADOV ZLIVANJA OBRAZOV

Magistrsko delo
Magistrski študijski program druge stopnje Elektrotehnika

Mentor: prof. dr. Vitomir Štruc
Somentor: asist. Marija Ivanovska Preskar

Ljubljana, 2025

UNIVERZA V LJUBLJANI
FE
Fakulteta za elektrotehniko

Tržaška cesta 25, p.p. 2999,
1000 Ljubljana, Slovenija
T: 01 476 84 11
dekanat@fe.uni-lj.si
www.fe.uni-lj.si

Št. teme: 00940 / 2025

Datum prijave: 3. 2. 2025

Univerza v Ljubljani, Fakulteta za elektrotehniko izdaja naslednjo nalogo:

Kandidat: **Nikola Marić**

Naslov: **Uporaba večmodalnih velikih jezikovnih modelov za zaznavanje napadov zlivanja obrazov**

Morphing attack detection using multimodal large language models

Vrsta naloge: Magistrsko delo
Magistrski študijski program druge stopnje Elektrotehnika

Tematika naloge:
Razišči problem napadov zlivanja obrazov, ki ogrožajo biometrične varnostne sisteme, ter preuči možnosti uporabe večmodalnih velikih jezikovnih modelov (MLLM) za njihovo zaznavanje. Analiziraj zmožnosti obstoječih odprtokodnih MLLM-ov brez dodatnega učenja in razvij strukturiran pristop k načrtovanju pozivov MLLM-jem za sistematično forenzično analizo. Z uporabo parametrično učinkovitih metod, kot je LoRA, prilagodi izbrane modele nalogi zaznavanja zlitih obrazov in jih ovrednoti na različnih podatkovnih zbirkah z uporabo standardne metodologije. Rezultate primerjaj z najsodobnejšimi metodami ter razpravljaj o prednostih, omejitvah in možnostih praktične uporabe razvite rešitve.

Ljubljana, 2. 9. 2025

prof. dr. Vitomir Štruc
Mentor

prof. dr. Igor Škrjanc
Predstojnik katedre

asist. Marija Ivanovska
Preskar
Somentor

prof. dr. Marko Topič
dekan

UNIVERZA V LJUBLJANI
FE
Fakulteta za elektrotehniko

Tržaška cesta 25, p.p. 2999,
1000 Ljubljana, Slovenia
T: +386 1 476 84 11
dekanat@fe.uni-lj.si
www.fe.uni-lj.si

Thesis number: 00940 / 2025

Date: 3. 2. 2025

University of Ljubljana, Faculty of Electrical Engineering issues the following thesis:

Candidate: **Nikola Marić**

Title: **Morphing attack detection using multimodal large language models**

Type of thesis: Master thesis

Topic of the thesis:
Investigate the problem of face morphing attacks in the context of biometric security systems and explore the use of multimodal large language models (MLLMs) for their detection. Evaluate the zero-shot capabilities of state-of-the-art open-source MLLMs and develop a structured prompt engineering framework for systematic forensic analysis. Apply parameter-efficient adaptation methods such as Low-Rank Adaptation (LoRA) to fine-tune selected models for morphing attack detection and validate their performance across multiple datasets using standard evaluation methodology. Compare the developed solutions with state-of-the-art approaches and discuss their advantages, limitations, and potential for real-world deployment.

Ljubljana, 2. 9. 2025

prof. dr. Vitomir Štruc
Mentor

prof. dr. Igor Škrjanc
Chair

asist. Marija Ivanovska
Preskar
Comentor

prof. dr. Marko Topič
dean

UNIVERZA V LJUBLJANI | FE Fakulteta za elektrotehniko

Tržaška cesta 25, p.p. 2999,
1000 Ljubljana, Slovenija
T: 01 476 84 11
dekanat@fe.uni-lj.si
www.fe.uni-lj.si

Spodaj podpisani študent, Nikola Marić, vpisna številka 64180424, avtor pisnega zaključnega dela študija z naslovom: Uporaba večmodalnih velikih jezikovnih modelov za zaznavanje napadov zlivanja obrazov,

IZJAVLJAM,

1. [1] a) da je pisno zaključno delo študija rezultat mojega samostojnega dela in da sem orodja umetne inteligence uporabljal odgovorno (predvsem s preverjanjem primarnih virov, brez vnašanja avtorsko zaščitenih del v orodja umetne inteligence, s kritičnim vrednotenjem rezultatov, ki so lahko pomanjkljivi, napačni ali neresnični) in sem izključno odgovoren za vsebino avtorskega dela, ki sem ga ustvaril;
   b) da je pisno zaključno delo študija rezultat lastnega dela več kandidatov in izpolnjuje pogoje, ki jih Statut UL določa za skupna zaključna dela študija ter je v zahtevanem deležu rezultat mojega samostojnega dela in da sem orodja umetne inteligence uporabljal odgovorno (predvsem s preverjanjem primarnih virov, brez vnašanja avtorsko zaščitenih del v orodja umetne inteligence, s kritičnim vrednotenjem rezultatov, ki so lahko pomanjkljivi, napačni ali neresnični) in sem izključno odgovoren za vsebino avtorskega dela, ki sem ga ustvaril;
2. da sem pridobil vsa potrebna dovoljenja za uporabo podatkov in avtorskih del v pisnem zaključnem delu študija in jih v pisnem zaključnem delu študija jasno označil;
3. da sem pri pripravi pisnega zaključnega dela študija ravnal v skladu z etičnimi načeli in, kjer je to potrebno, za raziskavo pridobil soglasje etične komisije;
4. da soglašam z uporabo elektronske oblike pisnega zaključnega dela študija za preverjanje podobnosti vsebine z drugimi deli s programsko opremo za preverjanje podobnosti vsebine, ki je povezana s študijskim informacijskim sistemom članice;
5. da na UL neodplačno, neizključno, prostorsko in časovno neomejeno prenašam pravico shranitve avtorskega dela v elektronski obliki, pravico reproduciranja ter pravico dajanja pisnega zaključnega dela študija na voljo javnosti na svetovnem spletu preko Repozitorija UL;
6. da dovoljujem objavo svojih osebnih podatkov, ki so navedeni v pisnem zaključnem delu študija in tej izjavi, skupaj z objavo pisnega zaključnega dela študija.
7. da dovoljujem uporabo mojega rojstnega datuma v zapisu COBISS.

V: Ljubljani
Datum: 2. 9. 2025

Podpis študenta: 

[1] Obkrožite varianto a) ali b).

University of Ljubljana

Faculty of Electrical Engineering

Nikola Marić

Morphing Attack Detection with Multimodal Large Language Models

Master thesis

2nd Cycle Postgraduate Study Programme Electrical Engineering

Supervisor: prof. dr. Vitomir Štruc

Co-supervisor: as. Marija Ivanovska Preskar

Ljubljana, 2025

Acknowledgments

I would like to express my sincere gratitude to all those who contributed to the successful completion of this master's thesis.

First and foremost, I am deeply grateful to my supervisor, Prof. dr. Vitomir Štruc, for his invaluable guidance, expertise, and support throughout this research. His insightful feedback, constructive criticism, and dedication to academic excellence have been instrumental in shaping the quality of this work.

I extend my special appreciation to my co-supervisor, As. Marija Ivanovska Preskar, whose unwavering commitment and collaboration have been fundamental to bringing this research to fruition. Her technical expertise in morphing attack detection, patient mentorship, and countless hours of discussion have not only enriched this thesis but also significantly contributed to my development as a researcher. Her willingness to share knowledge and provide detailed feedback at every stage of the project has been invaluable.

I am also deeply grateful to my friends and colleagues at the Jožef Stefan Institute (JSI) for fostering such a supportive and intellectually stimulating research environment. The readiness to share insights, engage in deep technical discussions, and provide constructive feedback was invaluable and made this work much better. I particularly appreciate their incredible patience and understanding with the extensive computational resources required for this project, especially my seemingly endless GPU usage for training and evaluating the large language models central to this research.

Finally, I acknowledge the computational infrastructure provided by JSI, without which the experiments presented in this thesis would not have been possible.

To all my friends, present, past, and beyond,
To all of those who weren't with us too long,
This one's for you!

Abstract

Face morphing attacks pose a significant threat to biometric security systems by enabling multiple individuals to authenticate with a single compromised credential i.e., a morphed face image. This thesis investigates the use of multimodal large language models (MLLMs) for morphing attack detection, demonstrating that foundation models trained on large-scale, heterogeneous data possess latent forensic capabilities that can be adapted for specialized security tasks.

We evaluate four open-source models in a zero-shot setting, including *Gemma-3 27B*, *Qwen2.5-VL 32B*, *Llama-4 Scout 17B*, and *Mistral Small 3.1 24B*, across diverse datasets covering landmark-based, GAN-based, and diffusion-based morphing attacks. Even without task-specific training, these models achieve measurable detection performance, confirming that multimodal language models inherently encode useful representations. To improve zero-shot detection reliability, we developed a structured forensic prompt, which guides the models through a systematic six-step procedure for detecting visual artifacts created during the blending of facial images. This structured prompting approach enhances both detection accuracy and interpretability of the outputs.

The primary contribution of the thesis lies in parameter-efficient fine-tuning through Low-Rank Adaptation (LoRA). Using only 0.61% of trainable parameters, we fine-tuned *Gemma-3 12B*. This fine-tuned model substantially outperformed its zero-shot counterpart, reducing the average Equal Error Rate by more than half. It achieved near-perfect detection on landmark-based morphs, competitive results on challenging GAN-based and diffusion-based morphs. Overall, this research establishes multimodal large language models as a viable and promising direction for morphing attack detection, combining generalization and interpretability with competitive performance against state-of-the-art approaches.

Key words: computer vision, deep learning, artificial intelligence, face analysis, morphing attack detection

Povzetek

Napadi zlitih obrazov (angl. face morphing attacks) predstavljajo resno grožnjo biometričnim varnostnim sistemom, saj omogočajo, da se z enim kompromitiranim poverilom, tj. obrazno sliko, overi več oseb. Čeprav obstoječe metode za zaznavanje napadov zlitih obrazov (MAD) kažejo obetavne rezultate,se soočajo z omejitvami pri posploševanju na različne tehnike zlivanja (angl. morphing techniques), pomanjkanja razložljivosti in odvisnosti od specializiranih učnih podatkov. V tej magistrski nalogi raziskujemo nov pristop, uporabo večmodalnih velikih jezikovnih modelov (MLLM) za zaznavanje napadov zlitih obrazov, pri čemer izhajamo iz domneve, da temeljni modeli, učeni na podatkih internetnega obsega (angl. internet-scale data), vsebujejo latentne forenzične analitične sposobnosti, ki jih je mogoče prilagoditi za specializirane varnostne naloge.

Predstavljamo evalvacijo odprtokodnih večmodalnih velikih jezikovnih modelov za zaznavanje zlitih obrazov, pri čemer obravnavamo štiri najsodobnejše modele: *Gemma-3 27B*, *Qwen2.5-VL 32B*, *Llama-4-Scout 17B* in *Mistral Small 3.1 24B*. Z obsežnimi eksperimenti na sedmih različnih podatkovnih zbirkah, ki vključujejo metode za generiranje zlitih obrazov na osnovi značilnih točk, GAN modelov, ter z difuzijskimi metodami, pokažemo, da ti modeli dosegajo merljive zmožnosti zaznavanja tudi brez dodatne optimizacije prednaučenega modela. Naš pristop brez optimizacije prednaučenega modela (angl. zero-shot) je pokazal, da je *Gemma-3 27B* najučinkovitejši model, saj je dosegel povprečno enako stopnjo napake (EER) 32,09 %, z izjemnimi rezultati pri difuzijskih napadih (6,15 % EER na podatkovni zbirki Greedy-DiM).

Za izboljšanje zaznavnih sposobnosti smo razvili strukturiran forenzični poziv (angl. prompt), ki večmodalne jezikovne modele vodi skozi sistematičen šeststopenjski postopek zaznavanja vizualnih artefaktov, nastalih kot posledica

zlivanja obrazov. V vsaki stopnji model oceni prisotnost značilnih anomalij in jim dodeli oceno zaupanja na lestvici od 0 do 10.000, pri čemer višja vrednost pomeni večjo gotovost v zaznavo artefakta. Ta pristop načrtovanje pozivov (angl. prompt engineering) je v povprečju izboljšal točnost zaznavanja za 10.3 %, ter zagotovil razložljive, strukturirane izpise, ki pojasnjujejo utemeljitev vsake odločitve, s čimer se večmodalni jezikovni modeli preoblikujejo iz "črnih škatel" v pregledna forenzična orodja.

Osrednji prispevek je uspešna prilagoditev splošnonamenskih večmodalnih jezikovnih modelov z učinkovitim učenjem parametrov. Z uporabo *Low-Rank Adaptation (LoRA)* in zgolj 0,61 % učljivih parametrov smo doučili model *Gemma-3 12B* na slikah sintetično zlitih obrazov, ki posnemajo napade zlitih obrazov. Rezultirajoči model je dosegel zelo dobre rezultate v različnih scenarijih ovrednotenja, v določenih kategorijah pa je celo presegel obstoječe najsodobnejše modele.

Model smo preizkusili na več podatkovnih zbirkah, ki zajemajo različne tipe morfiranih obrazov, in pokazali, da doučeni model *Gemma-3 12B-MAD* dosega konkurenčno učinkovitost v primerjavi z obstoječimi pristopi. Rezultati kažejo, da dosežemo nižje stopnje napak kot klasične metode zaznavanja zlitih obrazov, pri čemer se prednost še poveča pri strožjih operativnih pragovih, kjer močno zmanjšamo število lažnih zavrnitev. Posebej izstopa uspešnost pri zaznavanju zlitih obrazov narejenih na osnovi značilnih točk, kjer na podatkovni zbirki FRLL dosegamo skoraj popolne rezultate.

Ob primerjavi z metodami, ki se učijo na nenadzorovan ali samonadzorovan način, se naš pristop uvršča med najsodobnejše, saj zagotavlja stabilno in konkurenčno zaznavanje tudi v zahtevnejših primerih, kjer ohranja nizko stopnjo lažnih zavrnitev (BPCER). Poleg tega rezultati potrjujejo, da učinkovitost ne določa zgolj velikost uporabljenega modela, temveč predvsem ciljno učenje modela na specifično področje zaznavanja zlitih obrazov.

Z vidika praktične uporabe ponuja doučeni model pomembne prednosti. Sistem omogoča do 30-kratno pohitritev inferenciranja v primerjavi z pristopom brez optimizacije prednaučenega modela (iz 30 sekund na manj kot 1 sekundo na sliko). Ta drastična optimizacija je posledica uporabe manjšega modela (12B

namesto 27B), precej enostavnejšega poziva (angl. prompt), ki ne zahteva kompleksnega razmišljanja, in predvsem spremembe naloge iz generiranja besedila v klasifikacijo. Poleg tega sistem deluje na eni sami grafični kartici, kar ga naredi primernega za uporabo v realnih varnostnih okoljih.

Raziskava dokazuje, da so večmodalni veliki jezikovni modeli učinkovit pristop za zaznavanje napadov zlitih obrazov ter da jih je mogoče uspešno doučiti za specializirane biometrične varnostne naloge, pri čemer dosegajo primerljivo učinkovitost z najsodobnejšimi metodami. Delo odpira nove poti za izkoriščanje predhodnega učenja na podatkih internetnega obsega v domeno varnostnih aplikacij, ter ponuja okvir za prilagajanje večmodalnih jezikovnih modelov tudi za druge izzive zaznavanja biometričnih napadov.

Ključne besede: računalniški vid, globoko učenje, umetna inteligenca, analiza obrazov, napadi zlitih obrazov

Table of Contents

List of Figures

List of Tables

1 Introduction

Face-morphing attacks have emerged as a threat to biometric identification systems, exploiting the blending of facial images from two or more individuals to produce a synthetic face that can impersonate multiple identities, as illustrated in Figure 1.1 [2, 3]. By inserting a morphed face image into a document such as a passport, as demonstrated in Figure 1.2, an attacker and their accomplice may both pass automated face matching checks using the same credential [4]. These attacks undermine the integrity of security processes (e.g., border control) by allowing impostors, and potentially even criminals, to be authenticated as someone else, unless the morphing is detected. This risk has motivated intensive research into Morphing Attack Detection (MAD) techniques aimed at automatically distinguishing morphed images from bona fide ones [4, 5, 6, 1].

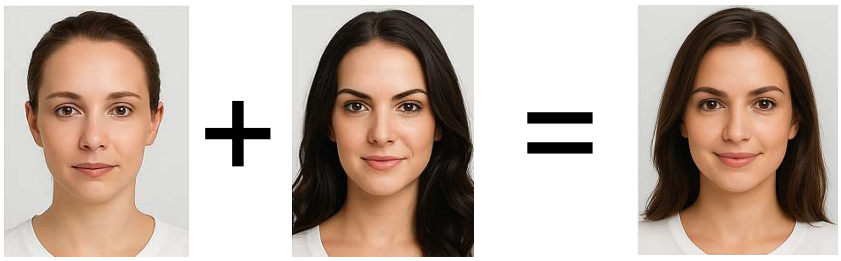


Figure 1.1: Face morphing process: Two or more source facial images (left) are blended to create a synthetic morphed face (right) that preserves recognizable characteristics of both contributing identities, enabling multiple individuals to authenticate using the same biometric credential.

MAD approaches can be categorized by their operational scenario. *Differential MAD* methods leverage a trusted reference (such as a live capture) alongside the suspect image, comparing two samples to discern inconsistencies [4, 7, 8].

Figure 1.2: Face morphing attack scenario: A morphed facial image embedded in an identity document (e.g., passport) allows both the primary holder and accomplice to successfully pass automated face recognition systems at border control or security checkpoints, compromising the integrity of biometric authentication.

Although effective in controlled settings such as passport issuance or automated border gates, differential methods are impractical when only a single image is available (for example, forensic analysis of a single photo) [4]. This has spurred the development of *single-image MAD* techniques that analyze one photo for telltale morphing artifacts [4, 5, 6].

Traditional single-image detectors often relied on hand-crafted visual features or simple classifiers, but these showed limited robustness. In many cases, studies yielded optimistic detection rates by evaluating on morphs generated with basic tools that leave visible artifacts [9]. Such conditions do not reflect sophisticated real-world morphs, which can be created with advanced software or generative models and contain far more subtle artifacts [9]. A key limitation of conventional MAD algorithms is their **poor generalizability**. Methods tend to perform well on the specific morphing techniques or datasets they were trained on, but can fail dramatically when confronted with morphs from unseen generation methods or novel attack scenarios [9, 6, 1].

Deep learning–based MAD systems have achieved higher accuracy than earlier approaches, yet they too are susceptible to overfitting to the training distribution [9, 6]. Moreover, most existing detectors operate as black-box classifiers, providing little or no explanation for their decisions, an issue in high-stakes ap-

plications where **interpretability is critical** [10]. These limitations highlight the need for more robust and transparent MAD solutions that can cope with the evolving landscape of morphing techniques.

Recent advances in artificial intelligence offer promising avenues to address these challenges. In particular, the integration of deep learning and image forensics has already improved morph detection capabilities, achieving significant gains in sensitivity and specificity over earlier methods. Researchers have explored a range of modern architectures, from convolutional neural networks (CNNs) and vision transformers to autoencoder-based anomaly detectors and generative models, to boost detection performance and generalization [5, 11]. For example, several works incorporate high-level CNN feature extractors or leverage residual noise analysis to distinguish morphed faces, achieving detection equal error rates well below 10 under evaluation conditions [12].

Notably, some studies have introduced domain-agnostic AI models into the MAD task. Foundation models trained on massive datasets, which demonstrate strong zero-shot learning abilities, are being repurposed for morphing attack detection [4, 10, 8]. The appeal of such models lies in their potential to combat the generalization problem. By virtue of learned broad visual representations, they can adapt to new attack types more readily than bespoke classifiers. For example, one recent approach adapts a pre-trained vision–language foundation model (*CLIP*) to the morph detection domain via lightweight fine-tuning, achieving performance on par with state-of-the-art dedicated MAD algorithms and even outperforming them in certain cross-dataset evaluations [4]. This suggests that **general-purpose vision models can be harnessed to build more resilient detectors**.

Beyond vision-first models, there is also growing interest in multimodal AI systems that combine visual and language understanding for security tasks. Leveraging the reasoning capabilities of large language models (LLMs) alongside image analysis could facilitate both high adaptability and interpretability in detection. In fact, early experiments have demonstrated that cutting-edge multimodal LLMs are capable of detecting face morphing attacks in a zero-shot manner, that is, without any task-specific training, while also producing human-readable explanations of how they identified the morph [10]. *Zhang et al. (2025)* [10]

report that a state-of-the-art multimodal LLM (*OpenAI's GPT-4 with vision*) showed remarkable generalization to previously unseen morphs, successfully flagging various morphing techniques and providing rationale for each decision [10]. This proof-of-concept highlights the untapped potential of advanced AI, including foundation models and multimodal LLMs, to significantly improve MAD by addressing both the robustness and explainability deficits of traditional methods.

In summary, face morphing attacks represent a serious security challenge for face recognition systems, and conventional detection methods encounter notable limitations in real-world deployment. The latest literature emphasizes a shift toward more generalized and intelligent solutions, from deep CNN-based detectors to adaptations of large pretrained models and even LLM-driven strategies, in order to keep pace with increasingly sophisticated morphing techniques [4, 1, 13]. Building upon this growing body of work, our present research investigates the use of open-source multimodal large language models for morphing attack detection. By exploring zero-shot and fine-tuning paradigms with state-of-the-art vision–language models, our aim is to advance the state of MAD toward greater accuracy, generalizability to unseen attacks, and improved transparency in decision-making. The following sections provide an overview of related work, describe our proposed MLLM-based MAD framework, and discuss experimental results in comparison to existing techniques.

1.1 Motivation

While recent preliminary investigations have demonstrated the potential of multimodal large language models (LLMs) for morphing attack detection (MAD) [10], significant research gaps remain that limit our understanding of their practical viability and optimal implementation strategies. These gaps provide the specific motivation for the comprehensive investigation presented in this thesis.

Current evaluations of multimodal LLMs for MAD have been limited in scope and methodology. Existing studies primarily focus on proprietary, closed-source models such as *GPT-4 Turbo* [14, 10], leaving the capabilities of open-source alternatives largely unexplored. This limitation creates a knowledge gap for re-

searchers and practitioners who require accessible, reproducible, and customizable solutions for morphing attack detection. Understanding the relative performance of open-source multimodal LLMs across different architectures and parameter scales is essential for establishing their practical utility in real-world security applications.

The absence of prompt engineering methodologies represents another significant limitation in current research. While early investigations have shown that appropriate prompting can elicit morphing detection capabilities from LLMs [10], these efforts lack principled frameworks for prompt design, optimization, and standardization. The development of structured analytical protocols that consistently guide multimodal LLMs through forensic examination processes remains an unexplored research direction with substantial implications for detection reliability and explanation quality in the field of morphing attack detection [15].

Furthermore, the potential for enhancing multimodal LLM performance through targeted fine-tuning has not been systematically investigated for morphing attack detection. Parameter-efficient adaptation techniques such as *Low-Rank Adaptation (LoRA)* [16] offer promising avenues for specializing foundation models to the MAD domain while preserving their generalization capabilities and computational efficiency [4]. However, the optimal strategies for fine-tuning multimodal LLMs on morphing detection tasks, including training data composition, architectural adaptations, and loss function design, remain unexplored.

Finally, the lack of comprehensive benchmarking against established MAD approaches prevents accurate assessment of where multimodal LLMs stand relative to current state-of-the-art methods. Rigorous comparative evaluations across diverse datasets and morphing techniques are necessary to establish the practical advantages and limitations of LLM-based approaches, informing their potential integration into existing security frameworks.

1.2 Goals of the Thesis

The primary objectives of this research are structured around four interconnected goals that collectively advance morphing attack detection through the application

of multimodal large language models (LLMs).

Evaluate Zero-Shot Morphing Attack Detection Capabilities of Open-Source Multimodal LLMs. The first goal involves conducting a comprehensive assessment of state-of-the-art open-source multimodal language models, including *Gemma-3 27B* [17], *Qwen2.5-VL 32B* [18], *Llama-4-Scout 17B* [19], and *Mistral Small 3.1 24B* [20], for their inherent ability to detect face morphing attacks without task-specific training. This evaluation encompasses testing across diverse morphing techniques and datasets to establish baseline performance and identify the most promising architectures for this security application.

Develop and Optimize Prompt Engineering Strategies for MAD. The second goal focuses on designing sophisticated prompt engineering methodologies that transform multimodal LLMs into effective forensic analysis tools for MAD. This involves creating structured analytical frameworks that guide models through systematic examination of facial images, incorporating domain-specific knowledge about morphing artifacts, and establishing standardized output formats that facilitate both automated evaluation and human interpretation.

Implement Parameter-Efficient Fine-Tuning for Enhanced Detection Performance. The third goal addresses the adaptation of multimodal large language models to the morphing attack detection domain through *Low-Rank Adaptation (LoRA)* fine-tuning techniques [16]. This includes developing training methodologies that leverage synthetic morphing attacks, comparing classification head approaches with generative fine-tuning strategies, and evaluating the impact of domain adaptation on both detection accuracy and generalization capabilities.

Establish Comprehensive Benchmarking Against State-of-the-Art Methods. The final goal involves conducting rigorous comparative evaluations against established MAD approaches, including supervised deep learning methods [5, 21], unsupervised anomaly detection techniques [6], and recent foundation model adaptations [4]. This benchmarking extends across multiple datasets repre-

senting various morphing generation techniques to provide assessments of relative performance, generalization capability, and practical deployment considerations.

1.3 Structure

The remainder of this thesis is organized into five chapters that systematically present the theoretical foundations, methodological approaches, experimental design, empirical results, and conclusions of this research.

Chapter 2 provides a comprehensive review of related work in morphing attack detection, establishing the theoretical and empirical context for our contributions. This chapter examines the evolution of face morphing techniques from traditional landmark-based approaches to sophisticated generative methods, surveys existing detection methodologies ranging from classical image forensics to modern deep learning approaches, and analyzes recent developments in foundation model applications for biometric security tasks.

Chapter 3 presents the detailed methodology underlying our experimental framework. This chapter describes the selection and configuration of multimodal large language models under investigation, outlines the systematic prompt engineering process developed for zero-shot morphing detection, details the parameter-efficient fine-tuning procedures using *Low-Rank Adaptation (LoRA)* [16], and specifies the comprehensive baseline methods implemented for comparative evaluation.

Chapter 4 describes the experimental design, dataset configuration employed throughout this research, and the empirical results obtained from both zero-shot and fine-tuned evaluations of multimodal LLMs for morphing attack detection. This chapter characterizes the diverse morphing attack datasets utilized for evaluation, including *FRLL-Morphs* [22], *FRGC-Morphs* [23], *FERET-Morphs* [24], *Greedy-DiM* [25], *MIPGAN-II* [26], *MorGAN* [27], *SMDD* [28], and *MorDiff* [29]. It further details the evaluation protocols and metrics applied for performance assessment, and presents the computational infrastructure and implementation considerations for reproducible experimentation. The analysis compares the performance of different prompt engineering strategies, evaluates the detection ca-

pabilities of various model architectures, assesses the effectiveness of fine-tuning approaches, and provides comprehensive benchmarking results against state-of-the-art detection methods across multiple datasets and morphing techniques.

Chapter 5 concludes the thesis by synthesizing the key findings, discussing their implications for the broader field of biometric security, acknowledging limitations of the current approach, and outlining promising directions for future research in foundation model applications for morphing attack detection and explainable biometric security systems.

2 Related Work

In this chapter, we examine different face morphing techniques, morphing attack detection (MAD) methods, and the use of foundational models and large language models (LLMs) for MAD up until now.

2.1 Face Morphing Techniques

Modern face morphing attacks fall into three broad families that differ in how the blend is constructed and in the artifacts they leave behind: *landmark-based* warping in the pixel domain, *GAN-based* generation in a model's latent space, and newer *diffusion-based* methods that synthesize a blend through iterative denoising. It is useful to distinguish *image-level* from *representation-level* morphs. Image-level (*landmark*) pipelines detect facial keypoints on two sources, establish correspondences, warp one or both faces to an intermediate geometry (e.g., piecewise-affine over Delaunay triangles), and linearly blend pixels—often with feathering, smoothing, or histogram adjustments to hide seams [12]. Typical cues are local ghosting, doubled edges, and texture discontinuities around the eyes, nostrils, lip borders, and hairlines. Representation-level methods instead embed both faces into a generator's latent manifold, combine them (by interpolation or identity-constrained optimization), then decode a synthetic face. These approaches usually remove visible seams but introduce generator-specific regularities—"frequency fingerprints"—that appear as structured spectral patterns or abnormal power distributions, even when the image looks flawless [1].

Landmark-based morphs are well represented in *FRLL-Morphs* under studio-quality, frontal conditions that make artifact analysis clean and consistent, as

(a) Face 01 (bona fide) (b) Face 01_10 (morph) (c) Face 10 (bona fide)

Figure 2.1: Landmark-based morphing (FRLL OpenCV): Facial keypoints are detected on source images (a, c), correspondences established, and faces warped to intermediate geometry using Delaunay triangulation, followed by pixel blending (b). Characteristic artifacts include ghosting, doubled edges, and texture discontinuities around eyes, lips, and hairlines.

demonstrated in Figure 2.1; FRLL also includes *OpenCV*, *FaceMorpher*, *AMSL*, and *StyleGAN2* subsets for cross-technique evaluation [22, 24, 23]. The *LMA-DRD* line of work extends landmark morphing to re-digitized (print–scan) scenarios that mimic passport issuance; the print–scan chain both attenuates and distorts landmark artifacts and adds its own noise, which is crucial for deployment realism [5]. Large synthetic corpora such as *SMDD* offer privacy-friendly training with landmark morphs at scale and are frequently used in competition settings like *SYN-MAD 2022* [28, 3]. Compared to the controlled FRLL setting, FRGC- and FERET-based morph sets bring more varied backgrounds, illumination, and legacy image characteristics; that domain shift alone can reduce S-MAD accuracy if a model has overfit to FRLL's studio conditions [23, 24].

GAN-based morphs illustrate representation-level generation, as shown in Figure 2.2. *MorGAN* pioneered latent-space blends using an encoder–decoder GAN and provides both GAN and landmark (LMA) attacks for comparison [27]. *StyleGAN2*-based pipelines underpin two important subsets: the StyleGAN2 splits inside *FRLL/FRGC/FERET* morph collections and *MIPGAN*, where identity-aware latent optimization produces high-resolution, ICAO-compliant morphs in both digital and print–scan form [26]. These images are typically photorealistic with few visible seams; their telltales are more often spectral/textural, consistent with the frequency-fingerprint perspective used in *SelfMAD* [1].

(a) Face 01 (bona fide) (b) Face 01_10 (morph) (c) Face 10 (bona fide)

Figure 2.2: GAN-based morphing (FRLL StyleGAN): Source faces (a, c) are embedded into StyleGAN's latent manifold, combined through latent space interpolation, then decoded as a synthetic face (b). These approaches introduce generator-specific regularities and structured spectral patterns.

Diffusion-based morphs are the latest wave, represented in Figure 2.3. Two strong variants are common in benchmarks: (i) diffusion autoencoder interpolation (e.g., *MorDIFF*), which linearly interpolates semantic latents and spherically interpolates stochastic components before decoding, and (ii) guided trajectory search/optimization (e.g., *Greedy-DiM*), which steers the denoising process using identity constraints to produce exceptionally seamless, high-match morphs; *Morph-PIPE* provides another diffusion pipeline used in recent evaluations [29, 25, 30]. These attacks preserve identity well and exhibit minimal pixel-space seams; when detectable, cues tend to be subtle periodicities or abnormal spectral densities, again aligning with the frequency-artifact view [1].

Because technique and acquisition matter, robust evaluation mixes both generation families and capture conditions. Under consistent studio imagery (*FRLL*), detectors can isolate technique effects; under more varied *FRGC*/*FERET* conditions, natural artifacts and background complexity stress generalization. Adding print–scan (*LMA-DRD*) exposes acquisition-chain shifts. Including GAN (*MorGAN*, *MIPGAN*) and diffusion (*MorDIFF*, *Greedy-DiM*, *Morph-PIPE*) sets probes representation-level fingerprints. The consistent lesson is that technique-agnostic S-MAD benefits from learning both families of cues: local pixel irregularities typical of landmark warping and global/frequency signatures typical of latent-space generation, thereby reducing cross-technique performance drops observed when training on only one morph type [1].

(a) Face 01 (bona fide) (b) Face 01_10 (morph) (c) Face 10 (bona fide)

Figure 2.3: Diffusion-based morphing (Greedy-DiM): Advanced synthesis through iterative denoising processes with greedy optimization strategies. Source faces (a, c) undergo diffusion-guided trajectory search to produce highly seamless morphs (b) with exceptional identity preservation while introducing subtle periodicities and abnormal spectral densities.

2.2 Morphing attack detection

Research in morphing attack detection distinguishes between *single-image (S-MAD)* detection and *differential (D-MAD)* detection using a trusted live image [9]. Single-image methods operate on just the document or uploaded photo, which is the most common and challenging scenario. Differential methods, in contrast, have the benefit of a second image of the same person (e.g., a live capture at border control) for comparison [9]. Most recent research (and all the methods discussed below) concentrate on single-image detection *(S-MAD)*.

2.3 Hand-Crafted Morphing Attack Detection (MAD)

Initial efforts to detect morphed faces borrowed techniques from image forensics and facial biometrics. Researchers experimented with handcrafted features that could expose the subtle artifacts of blending. Texture-based detectors were among the first: for example, filtering the image with *Local Binary Patterns (LBP)* [31] or *Binarized Statistical Image Features (BSIF)* [32] and training an SVM classifier [33, 9]. The rationale is that morphing disrupts the natural micro-texture of genuine face images. Simple LBP-based classifiers performed reasonably on

known morph datasets (often attaining >90% TPR at 10% FPR in intra-dataset tests [9]) and even showed that fusing multiple texture features (LBP, SIFT, HOG, etc.) could boost accuracy [9].

Another line of attack used noise analysis, operating on the assumption that morphing two images inconsistently alters the sensor noise patterns. Methods based on *Photo Response Non-Uniformity (PRNU)* analysis attempted to detect morphs via inconsistencies in the camera noise fingerprint [34, 35]. Similarly, frequency-domain cues (e.g., artifacts revealed by Fourier or wavelet transforms) and edge inconsistencies were investigated. These classical approaches demonstrated that morphed images are not "perfectly natural" and often carry detectable traces.

However, they fell short in generalization. As *Scherhag et al. (2019)* [36] note in their survey, detectors relying on general image descriptors suffered severe performance drops when evaluated on morphs from new sources or algorithms [36]. For instance, an LBP-based detector trained on landmark morphs might misclassify most GAN-morphs as bona fide, and vice versa. Overall, the pre-deep learning era established useful baselines (often with Equal Error Rates in the 10–20% range on seen attacks [9]) but underscored the need for deep learning based methods that can automatically discover more discriminative morph cues.

2.4 Deep Learning-Based Morphing Attack Detection (MAD)

Research in MAD has progressed from simple forensic analysis to advanced deep learning methods. We organize current approaches into several categories and highlight representative methods in each, including their contributions, evaluation results, and limitations.

2.4.1 Supervised Deep Learning-Based MAD

Supervised CNN detectors remain the dominant paradigm through 2020–2023, typically trained on labeled bona fide versus morphed images. Many architectures have been explored, from standard classification backbones (*VGG* [37], *ResNet* [38], *Inception* [39]) to bespoke networks for morph detection.

A notable method, *MixFaceNet-MAD*, originates from the efficient face recognition architecture *MixFaceNet* (*Boutros et al., 2021* [21]), recognized for its compact size and robust performance. *Damer et al.* adapted MixFaceNet for morph detection [28], training it from scratch on the synthetic *SMDD* morph dataset and evaluating its performance on real-world data. Despite its lightweight design, MixFaceNet-MAD demonstrated commendable cross-dataset accuracy. For instance, when trained on *SMDD* and evaluated on datasets like *FRLL*, *FERET*, and *FRGC*, MixFaceNet-MAD established a baseline performance level for comparison. However, its supervised nature necessitates extensive and diverse training datasets to effectively handle various morph styles. Evaluation within the SPL-MAD study indicated competitive Equal Error Rates (EER) on certain morph types but highlighted limitations in detecting GAN-based morphs, suggesting the necessity of specialized training techniques or augmentation strategies to enhance generalization [6].

Several MAD approaches also leverage variants of widely-used architectures such as *Inception* [39] and *ResNet* [38]. One prominent baseline is *InceptionV3-MAD*, as discussed in recent evaluations [5]. While these models typically achieve strong performance on intra-dataset tests (often obtaining APCER and BPCER below 5% on known attacks), they tend to overfit, resulting in poorer cross-dataset generalization [6]. For instance, SPL-MAD experiments demonstrated significantly higher errors for Inception-based models on unseen morph types (e.g., FRLL-OpenCV morphs exhibited a BPCER@1% of approximately 24.32%), underscoring challenges in generalization. These findings suggest that, without specific measures, standard deep CNNs inherently learn dataset-specific cues, limiting their robustness to novel morph attacks. However, when augmented with specialized training or attention mechanisms, Inception- and ResNet-based MAD models provide foundational structures for more advanced detection methods.

Pixel-Wise MAD (PW-MAD), proposed by *Damer et al. (2021)* [5], offers another compelling supervised approach. PW-MAD employs pixel-wise binary supervision by using annotated masks highlighting morph artifacts versus genuine regions. These masks, approximated from source-image differences, guide CNNs to predict pixel-level morph regions, thereby fostering a richer understanding of morphing artifacts. This approach demonstrated superior generalization capabilities compared to conventional single-label classifiers, notably enhancing detection performance against unseen morph attacks. Damer et al.'s experiments indicated significant improvements in detecting novel morphs due to the explicit pixel-level supervision provided by PW-MAD [5].

2.4.2 Unsupervised and Self-Supervised Approaches for Generalized MAD

While powerful, supervised MAD methods historically suffered from poor generalization, excelling on known attack types but faltering on novel morphs. Recent efforts have aimed to improve robustness through network architecture innovations, data augmentation, and training paradigms such as self-supervision or one-class learning. Several notable unsupervised and self-supervised deep learning-based MAD approaches have emerged in this context.

Fang et al. (2022) introduced *Self-Paced Learning MAD (SPL-MAD)*, an unsupervised method leveraging abundant face recognition data, which may inadvertently include morphs, treating MAD as an anomaly detection problem [6]. SPL-MAD employs a *Convolutional Autoencoder (CAE)*, trained solely on bona fide images, using reconstruction error to detect morphs. Crucially, they incorporated a self-paced curriculum: initially training on "easy" samples, gradually integrating more challenging ones while automatically weighting samples by reconstruction loss. This approach ensures outliers do not skew the training, resulting in impressive generalization and outperforming many supervised methods across diverse datasets, including benchmarks like *MorPH* and *LMA* [6]. The primary strength of SPL-MAD lies in its open-set detection capability, not presuming knowledge of specific attack styles. However, it requires meticulous training on ideally morph-free datasets, as hidden morphs could undermine its effectiveness. Additionally, subtle morphs resembling bona fide faces might evade detection.

Despite these limitations, SPL-MAD marked a significant step towards general solutions and set the groundwork for diffusion-based methods.

Building upon this foundation, *Ivanovska and Štruc (2023)* proposed *MAD-DDPM*, a one-class approach utilizing *Denoising Diffusion Probabilistic Models (DDPMs)* [40]. MAD-DDPM trains a diffusion autoencoder exclusively on bona fide images, capturing their manifold. During inference, high reconstruction error or low likelihood indicates morph images as out-of-distribution. Evaluations on *CASIA-WebFace* (training) and *FRLL-Morphs*, *FERET-Morphs*, and *FRGC-Morphs* (testing) showed that MAD-DDPM achieved competitive or superior detection rates compared to supervised models, particularly excelling at detecting diffusion-based morphs (e.g., *MorDIFF*), leveraging its inherent diffusion modeling strength [40]. Despite outstanding performance, its high computational cost and susceptibility to falsely flagging other anomalies (e.g., heavy makeup or plastic surgery) remain limitations. Nevertheless, MAD-DDPM currently represents state-of-the-art unsupervised MAD, significantly lowering Equal Error Rates (EERs).

Beyond pure one-class approaches, some works have pursued disentanglement and self-supervision to improve generalization. *Neto et al. (2022)* introduced *OrthoMAD*, emphasizing orthogonal identity disentanglement through a novel regularization term integrated into a ResNet-18 classifier [11]. OrthoMAD produces two latent vectors per face image, ideally encoding identical information for bona fide images, thus making orthogonality challenging. Conversely, morphs containing multiple identities are encouraged to generate distinct orthogonal vectors. Tested across five morph types in *FRLL* datasets, OrthoMAD achieved superior APCER/BPCER performance, especially for *StyleGAN2* and *OpenCV* morphs, benefiting from simplicity and computational efficiency compared to autoencoder-based approaches. However, its orthogonality assumption might falter with highly skewed morph blends, and care must be taken to avoid false positives in bona fide images. Nevertheless, OrthoMAD effectively integrates face recognition knowledge directly into MAD tasks, providing an accessible and practical solution.

Similarly focusing on identity disentanglement, *Caldeira et al. (2023)* introduced *IDistill*, an interpretable method explicitly separating identities present in morphs using a two-part model [13]. An initial autoencoder separates identity

features into distinct latent vectors, subsequently distilled into a classifier network determining morph presence based on identity vector distances or orthogonality. IDistill achieved state-of-the-art results on multiple datasets, outperforming or remaining competitive with prior methods. Its interpretability allows potential identification of individuals in morphs, progressing towards demorphing or accomplice identification. However, IDistill demands a robust face recognition backbone and may struggle with balanced morph blends. Despite these challenges, its balanced performance and interpretability, evaluated on public datasets and openly available for reproducibility, highlight a meaningful advancement in MAD research.

Lastly, *Ivanovska et al. (2025)* proposed *SelfMAD*, a self-supervised framework generating simulated morph-like artifacts during training, negating the need for actual morph supervision [1]. This approach encourages learning generic morph cues rather than specific attack types, dramatically improving generalization. SelfMAD achieved groundbreaking results, significantly reducing detection errors (EER) by over 64% compared to previous unsupervised methods and by 66% versus leading supervised methods in cross-dataset scenarios. The success of SelfMAD underscores the effectiveness of leveraging unlabeled data and synthetic perturbations, aligning with contemporary trends emphasizing methods resilient to evolving morphing techniques without exhaustive morph galleries.

Apart from the mentioned approaches, one thread of MAD research leverages the observation that morphing processes often degrade certain image quality measures. By treating morph detection as an image quality assessment problem, these methods avoid explicit supervised training on morphs; instead, they exploit *Face Image Quality Assessment (FIQA)* [41, 42, 43, 44] or general *Image Quality Assessment (IQA)* [45, 46, 47, 48, 49, 50] scores to distinguish morphs from bona fide images. *Fu and Damer (2022)* conducted an extensive study in this direction [51], examining several unsupervised IQA-based detection techniques.

Among the explored methods is *MagFace*, a face recognition model developed by *Meng et al. (2021)* that inherently provides a face quality score based on the magnitude of the embedding alongside its identity embeddings [41]. In the context of MAD, MagFace's quality scores tend to be lower for morphed images than for genuine ones [51]. Intuitively, a morph, representing an averaged

or blended face, is less tightly clustered around any single identity, resulting in embeddings with lower utility for recognition, which MagFace captures through reduced embedding magnitudes. Fu and Damer found MagFace scores provided effective separability between bona fide images and morph attacks across various datasets and morph types [51]. For instance, on the *MorGAN* dataset, MagFace achieved Attack Presentation Classification Error Rates (APCER) of approximately 0.61–0.66 for both GAN-based and landmark-based morphs, despite these being relatively high error rates in absolute terms [51]. On the *LMA-DRD* print-scan set, MagFace similarly outperformed other quality metrics, indicating robust detection even after image degradation [51]. Overall, MagFace as an unsupervised detector achieved accuracy exceeding 70% in distinguishing morphs within mixed-dataset evaluations [51]. The primary strength of MagFace lies in its generalization ability, given that it was not explicitly trained for morph detection, yet successfully generalizes quality assessments to unseen morphing attacks, particularly GAN-based morphs. However, a limitation arises with subtler landmark-based morphs, which occasionally preserve higher facial utility and may evade detection.

Another promising approach explored by Fu and Damer is *CNNIQA*, a no-reference IQA CNN originally proposed by *Kang et al. (2014)* [45], which predicts quality scores based on local image patches [45]. In morphing attack detection, inverted IQA scores (i.e., interpreting lower quality as indicative of a morph) effectively signal morphing artifacts such as ghosting or double edges in facial features [51]. CNNIQA proved highly sensitive to local distortions, making it particularly suitable for detecting subtle inconsistencies introduced by morphing processes. Fu and Damer demonstrated that CNNIQA performed exceptionally well on landmark-based morphs, thereby complementing the capabilities of MagFace [51]. Specifically, on datasets like *FRLL-Morphs* and *FERET/FRGC* morphs, which predominantly feature landmark-based morphs, CNNIQA achieved among the lowest Average Classification Error Rates (ACER), frequently below 0.30 at a threshold of 20% Bona Fide Presentation Classification Error Rate (BPCER) [51]. The combination of MagFace and CNNIQA was suggested as an effective fusion approach to comprehensively detect both GAN-based and landmark-based morph attacks [51].

CNNIQA's primary advantage lies in its independence from face labels and

morph-specific training, solely relying on evaluating the naturalness of image quality. However, this method has a notable limitation in its potential to misclassify genuine images affected by benign degradations such as poor lighting or compression artifacts, interpreting them erroneously as morph traces. Thus, calibration of IQA-based methods may be necessary to mitigate false positives stemming from genuine images of lower quality. Nonetheless, unsupervised quality metrics like MagFace and CNNIQA have demonstrated generalized detection accuracies exceeding 70%, supporting the concept that morphing processes inherently leave detectable quality footprints. Such methods offer baseline detection capabilities without requiring morph-specific training data, proving valuable in real-world scenarios where novel morphing techniques continually emerge.

2.4.3 Foundation Models and Multimodal MAD

As the face morphing arms race continues, researchers have begun exploring large pre-trained foundation models and multimodal AI to push the envelope in MAD. Foundation models like *CLIP* (Contrastive Language–Image Pre-training) encode extremely rich, generic visual features by training on huge image datasets with natural language supervision. Such models demonstrate remarkable zero-shot generalization to new tasks and domains [4].

Caldeira et al. (2025) [4] recognized this potential and proposed *MADation*, the first adaptation of a vision–language foundation model for morph detection [4]. Their approach fine-tunes CLIP's image encoder with lightweight *LoRA* (Low-Rank Adaptation) layers and a simple classification head for MAD, rather than training a CNN from scratch [4]. The intuition is that CLIP's broad visual knowledge (e.g., understanding of faces, textures, anomalies) confers a strong starting point, which can be adapted to detect morph-specific anomalies with relatively few parameters. Indeed, MADation achieved competitive results with state-of-the-art detectors, even surpassing them in some cross-evaluation scenarios [4]. This is notable because it suggests that **a small amount of task-specific fine-tuning on a foundation model can yield generalization comparable to specialized methods**.

Alongside vision foundation models, researchers have also looked at multi-

modal large language models (LLMs) for MAD. *Zhang et al. (2025)* [10] presented a zero-shot approach using *GPT-4 Vision*, essentially prompting a multimodal LLM to act as a morph detector by providing image inputs and questions [10]. Surprisingly, they found that a powerful LLM with vision capability (without any fine-tuning on morph data) could achieve respectable accuracy in distinguishing morphs, while also providing a textual explanation of its decision [10]. This "AI judge" approach highlights the incredible generalization of foundation models: a system like *ChatGPT* (vision-enabled) has implicitly learned about face realism versus artifacts through its vast training data [10]. As a result, it can detect anomalies in a face image that correlate with morphing, despite never being trained for that task, and even articulate the reasoning (e.g., "the image has two sets of eyebrows, indicating a blend"). Such capabilities combine robustness and explainability, two facets that traditional MAD solutions often lack [10].

In summary, the field of Morphing Attack Detection (MAD) has evolved significantly, moving from initial methods based on handcrafted features and image quality analysis to more sophisticated deep learning techniques. Supervised models, such as *MixFaceNet-MAD* [21], and *PW-MAD* [5], demonstrated strong performance but often struggled with generalization, overfitting to the specific morphing techniques they were trained on. This led to the development of unsupervised and self-supervised approaches like *SPL-MAD* [6] and *SelfMAD* [1], which improved robustness by learning to detect anomalies without requiring labeled morph examples.

More recently, the focus has shifted towards leveraging large-scale, pre-trained foundation models. Approaches like *MADation* [4], which adapts the *CLIP* vision-language model, and preliminary studies using *GPT-4* [10], have shown that these generalist models possess latent forensic capabilities. They have achieved competitive results, suggesting a new paradigm where broad visual knowledge can be adapted for specialized security tasks, addressing the critical challenge of generalization against new and unseen morphing attacks.

However, even these state-of-the-art detectors face an ongoing arms race, where their accuracy is challenged by every new morphing technique, such as those based on advanced GANs [26, 52] and diffusion models [29, 25]. Furthermore, early research into foundation models for MAD has often relied on

closed-source systems or narrowly scoped studies.

This master's thesis builds directly on these pioneering efforts, positioning itself at the forefront of foundation-model-based MAD. We address the limitations of prior work by systematically evaluating open-source Multimodal Large Language Models (MLLMs) for morph detection. Our contribution extends the current paradigm in three key ways: first, by developing a structured prompt engineering framework to unlock and guide the inherent forensic abilities of these models; second, by exploring parameter-efficient fine-tuning (*LoRA*) [16] to specialize these general-purpose models for morph detection; and third, by providing a comprehensive benchmark against traditional and state-of-the-art methods. In doing so, this work aims to advance morph detection beyond prior studies by enhancing accuracy, ensuring interpretability, and improving resilience to unknown attacks through the adaptable power of MLLMs.

3 Methodology

3.1 Approach Overview

Our study comprises two major experimental pipelines: (i) zero-shot prompting, where we evaluate off-the-shelf multimodal LLMs on the morph detection task without additional training, and (ii) LoRA fine-tuning, where we adapt one model on a custom training set and then evaluate it. In both cases, we conduct image-only single-image morphing attack detection (S-MAD), i.e., the detector analyzes each image in isolation, with no reference image for comparison. This setting reflects the practical scenario of detecting a morph from a single submitted photo, and it is more challenging than differential MAD (which has a live reference capture).

The zero-shot pipeline involves designing a prompt, that would be appropriate for all evaluated models (see Section 3.3) and capturing their raw textual outputs on a suite of test images, followed by post-processing and metric computation. The fine-tuning pipeline involves training lightweight *LoRA* adapters on a balanced morph/non-morph dataset (see Section 3.4) and then evaluating the tuned model. All steps operate on single facial images only, emphasizing that our system does not rely on any paired or differential inputs.

3.2 Multimodal LLMs

We evaluate four multimodal large language models (LLMs) that accept images as input. All are open or research-access models smaller than GPT-4, chosen to

represent the current state-of-the-art in vision-language foundation models and to explore a range of model sizes.

Gemma-3 27B Vision [53]. A 27-billion-parameter multimodal model released by Google DeepMind (Apache 2.0 license) [54]. Gemma-3 uses a ViT-L (Vision Transformer Large) as its image encoder, fused with a Mixture-of-Experts text transformer. It supports a long context (up to 128k tokens) and is instruction-tuned for both image and text inputs. We included Gemma-3 for its state-of-the-art performance on vision-language tasks and because its open checkpoint allows efficient fine-tuning with LoRA. (We used the `google/gemma-3-27b-it` checkpoint from HuggingFace, 2025-01 version [17], for all experiments.)

Qwen2.5-VL 32B [55]. A 32-billion-parameter vision-language model in the Qwen series from Alibaba, licensed under Apache 2.0. The version used is `Qwen/Qwen2.5-VL-32B-Instruct` from HuggingFace [18], featuring a native dynamic-resolution ViT trained from scratch, window attention, and dynamic resolution processing to efficiently handle images and long-form video inputs with second-level temporal localization and precise object grounding via bounding boxes or points [55]. This instruction-tuned "chat" variant supports multi-round image-based dialogue and excels at fine-grained tasks such as OCR, document parsing, chart/table comprehension, and diagram understanding. It represents a flagship scale model in the series and demonstrates state-of-the-art performance on multimodal comprehension benchmarks, particularly in visual question answering and structured data extraction [55].

Llama-4-Scout 17B [19]. A multimodal model from Meta's Llama 4 family (Llama Community License), featuring 17 billion active parameters with 16 mixture-of-experts, totaling 109 billion parameters, and offering an industry-leading 10 million token context window for long-form vision-language reasoning. "Scout" is trained from scratch, supports native early-fusion multimodal inputs, and excels in grounded image understanding, captioning, visual question answering, and document comprehension across 12 languages.

We included a quantized variant of Llama-4-Scout to represent a high-efficiency mid-size multimodal LLM, evaluating whether such a compact, quantized model can perform well in morphing-attack detection. (Checkpoint: `meta-llama/Llama-4-Scout-17B-16E-Instruct`, released April 5, 2025; inference deployment via vLLM quantization for accelerated execution.)

Mistral Small 3.1 24B [20, 56]. A 24-billion-parameter multimodal model in the Mistral Small 3.1 series from Mistral AI, released under the Apache 2.0 license (March 17, 2025). This instruction-tuned variant (`mistralai/Mistral-Small-3.1-24B-Instruct-2503`) supports both text and vision inputs and extends context length up to 128k tokens while maintaining top-tier text performance [20]. It excels at multimodal reasoning, document and image understanding, multilingual tasks across dozens of languages, agent-like function calling, and fast inference even on a single RTX 4090 or a 32 GB-RAM Mac. On multimodal instruct benchmarks (e.g., ChartQA, DocVQA, MM-MT-Bench), it delivers performance on par with or exceeding larger models like Gemma 3, with strong results across vision-heavy tasks such as OCR and visual question answering (MMMU: $\sim$ 64%, ChartQA $\sim$ 86%, DocVQA $\sim$ 94%) [56]. We included this model to explore whether a moderately-sized open model can match larger architectures on visual reasoning and structured data interpretation, while offering ultra-efficient deployment. (Checkpoint: `mistralai/Mistral-Small-3.1-24B-Instruct-2503`, March 2025 release.)

Summary. The four multimodal LLMs selected for evaluation: *Gemma-3*, *Qwen2.5-VL*, *Llama-4-Scout*, and *Mistral Small 3.1*, span a diverse spectrum of architectures, parameter scales, inference efficiency, and licensing openness. This selection allows us to comprehensively assess the current capabilities and limitations of state-of-the-art multimodal models in morphing attack detection tasks. By including models that vary in size from moderately sized to large, as well as exploring quantized inference and long-context support, our evaluation provides valuable insights into both performance trade-offs and practical deployment considerations.

3.3 Prompt Engineering for Zero-Shot Morph Detection

The efficacy of a Multimodal Large Language Model (M-LLM) in a specialized, zero-shot task like facial morph detection is profoundly dependent on the quality of its prompt. The prompt serves not merely as a question but as a complex instruction set that guides the model's analytical focus, reasoning process, and output format. Recognizing this, we undertook a systematic prompt engineering methodology to develop a framework capable of eliciting precise and interpretable forensic analysis from M-LLMs. Our process evolved from simple classification queries to a sophisticated, structured analysis prompt designed to quantify suspicion rather than merely classifying an image.

3.3.1 Prompt Development and Discovery Process

The development of effective prompts for morphing attack detection followed an iterative approach driven by performance evaluation and computational constraints. Due to the substantial inference times required for multimodal LLM evaluation (30 seconds per image on average), initial prompt development was performed on a representative subset of data before full-scale evaluation was performed with promising configurations.

We started by testing *OpenCV* and *StyleGAN* morphs from the *FRLL* dataset [22], chosen to cover two different cases, as illustrated in Figure 3.1. *OpenCV* morphs show more visible artifacts and are therefore easier to detect; even with "bad" prompts they often get flagged, which lets us track improvements in prompting even when the harder cases sit near an EER of 50%. *StyleGAN* morphs, being GAN-based and more sophisticated, were chosen as an indicator of the prompt's potential in a cross-dataset setting. We tested our prompt variants on each LLM to find those suitable for the final evaluation, since some prompts worked well on certain models but produced incoherent or unusable responses on others.

Inspired by prior work in M-LLM prompting for analytical tasks [57] and Chain-of-Thought (CoT) reasoning [58], we transitioned to more elaborate

(a) FRLL OpenCV

(b) FRLL StyleGAN

Figure 3.1: Representative examples from the two morphing techniques used for prompt development testing: (a) OpenCV morphs and (b) StyleGAN morphs.

prompts. The core hypothesis was that by instructing the model to perform a series of analytical steps, we could encourage a more robust reasoning process, helping it identify subtle inconsistencies that a high-level classification might miss.

This led to the introduction of domain-specific knowledge into the prompts. We incorporated explicit guidance on which visual artifacts to look for, drawing from established literature on digital image forensics and morphing attack characteristics [1]. For instance, prompts were enriched with instructions like: "Scrutinize the image for common morphing artifacts, such as ghosting around the eyes, unnatural skin texture, and blurred contours at the hairline."

3.3.2 Development of an Integrated Scoring and Analysis Framework

Our preliminary prompting tests (see Section 5.1.1) indicated a key limitation in using Multimodal Large Language Models (MLLMs) for classification: the models struggled to accurately map their visual observations to discrete binary labels such as "morph" or "bonafide". To address this and capture more nuanced model judgments, a numerical confidence scoring system was developed and implemented. The initial iteration of this system utilized a 0–100 scale with a required precision of two decimal places.

However, analysis of the resulting score distributions revealed significant deficiencies in this initial approach. The models exhibited a strong tendency to

disregard the specified decimal precision, consistently generating round-number scores (e.g., 10.00, 20.00, 50.00, 80.00). This clustering behavior demonstrated that the 0–100 scale lacked the necessary resolution to elicit fine-grained confidence assessments. The outputs were interpreted as arbitrary scores rather than genuine confidence judgments, necessitating a methodological refinement.

To overcome these limitations, a comprehensive, multi-pronged strategy was engineered. This refined methodology involved the concurrent enhancement of the scoring mechanism, the formulation of a structured analytical process, and the enforcement of a standardized output format.

The core of this strategy was the expansion of the confidence range to 0–10,000. The increased granularity of this scale was designed to eliminate the previously observed score clustering and enable the model to express more precise degrees of confidence. To complement this, a semantic mapping framework was introduced to provide a clear interpretative guide for the numerical scores. This framework establishes a direct link between the quantitative output and a qualitative assessment by providing both simplified and detailed semantic descriptions, ranging from basic threshold definitions to comprehensive interpretative guidelines for forensic analysis.

Example of Semantic Scoring Guides.

A. Simplified Semantic Guide

- 0–1,000: Strong evidence of authentic face.
- 1,000–3,000: Likely authentic with minor irregularities.
- 4,000–7,000: Uncertain, requires careful analysis.
- 7,000–9,000: Likely morphed with moderate evidence.
- 9,000–10,000: Strong evidence of morphing attack.

B. Detailed Interpretative Guide (Excerpt)

- Score 9,000–10,000 (Very High / Near Certainty): Overwhelming and clear evidence of morphing. Multiple, strong artifacts are easily identifiable and create an incoherent image.
- Score 1,000–4,000 (Low to Moderate Suspicion): One or two minor, inconclusive artifacts are present (e.g., slight unnatural smoothness, minor asymmetry). These could potentially be explained by compression, lighting, or natural features, but warrant a degree of suspicion.

The culmination of our design process is a comprehensive prompt that integrates guided reasoning, fine-grained scoring, and a structured output format. This final prompt, used for our primary analysis, is designed to act as a complete forensic analysis protocol for the M-LLM.

In tandem with these scoring enhancements, a structured, six-step analytical framework was formulated to guide the models through a systematic forensic examination. This framework ensured comprehensive analysis by directing the models to sequentially evaluate core facial features, facial geometry, skin texture, boundary characteristics, lighting consistency, and overall identity coherence.

- **Step 1: Core Facial Feature Analysis:** Focuses on high-information areas like eyes and lips.
- **Step 2: Facial Geometry and Symmetry Analysis:** Guides attention to structural coherence.
- **Step 3: Skin Texture and Detail Analysis:** Probes for unnatural smoothing or lack of detail.
- **Step 4: Boundary and Edge Analysis:** Checks for common blending artifacts at the face perimeter.
- **Step 5: Lighting and Color Consistency Analysis:** Examines the physical plausibility of the image.

- **Step 6: Identity Coherence Analysis:** A holistic check for whether the features believably belong to a single individual.

The initial results showed significant performance improvements, which can be attributed to providing the LLMs with a much clearer task and richer context. Experimental validation revealed strong interdependencies between analytical components. Systematic removal of individual steps, even those appearing to contribute minimally, resulted in consistent performance degradation. This finding established the necessity of comprehensive rather than selective analysis, with each step contributing to overall detection robustness.

Given the complexity of the multi-step prompts, enforcing a strict output format became necessary to ensure the model provided coherent, readable, and machine-parseable results. This standardization was crucial for the systematic evaluation of the detailed analytical outputs generated by the framework.

3.3.3 Final Prompt Variants

The iterative development process yielded three main prompt configurations representing key methodological milestones:

Prompt 1 **(Structured Forensic Analysis – Semantic Guide A).** Complete six-step protocol with simplified semantic scoring and basic threshold definitions. Establishes foundational structured approach.

Prompt 2 **(Extended Forensic Analysis – Semantic Guide A).** Augmented version incorporating additional sub-questions and expanded contextual instructions. Tests impact of increased complexity.

Prompt 3 **(Optimized Forensic Analysis – Semantic Guide B).** Refined implementation featuring detailed interpretative guide, streamlined instructions, and mandatory reasoning explanations. Represents optimal balance of structure and efficiency.

The complete text of all three prompts is provided in Appendices A.1–A.3.

3.4 Gemma-3 12B Fine-Tuning

We fine-tuned the *Gemma-3 12B Vision* model using LoRA (Low-Rank Adaptation) [16] adapters to specialize it for morph detection. LoRA allows us to inject a small number of trainable parameters into the frozen pre-trained model, greatly reducing the memory and data needs for fine-tuning [16]. We chose *Gemma-3* for fine-tuning due to its strong zero-shot performance and open model weights.

3.4.1 Self-Supervised attack simulation pipeline

The methodology for constructing the training and validation datasets is central to this work and is adopted directly from the self-supervised approach detailed in the SelfMAD paper [1]. This method allows the model to learn generalizable features of morphing attacks without being trained on any real morphed images, thereby preventing overfitting to the artifacts of specific morphing techniques. The entire process relies on using only bona fide images and synthetically introducing a wide range of artifacts that mimic those found in real-world morphing attacks, as demonstrated in Figure 3.2.

To create a robust training environment, the bona fide images were processed through a multi-stage pipeline designed to simulate attacks. This process begins by first augmenting each bona fide image (I_{OS}) to simulate subtle, real-world variations in lighting, color, and quality. This step, which produces an augmented image (I_{AS}), ensures that the model becomes robust to normal photographic variations. Both the original and augmented images (I_{OS} and I_{AS}) serve as the "bona fide" class (label 0) during training.

The pipeline then introduces artifacts that mimic morphing attacks. To replicate the ghosting, blurring, and geometric misalignments characteristic of traditional, landmark-based techniques, the augmented image (I_{AS}) undergoes geometric transformations (e.g., elastic deformation) and is then blended with the original, non-augmented image (I_{OS}). This blending is guided by a binary mask

(a) Original bona fide image

(b) Synthetically generated attack

Figure 3.2: Self-supervised attack simulation example: (a) Original bona fide image from the training dataset, and (b) the corresponding synthetically generated attack created through the SelfMAD pipeline, incorporating pixel-level artifacts (geometric transformations, blending) to simulate morphing attack characteristics without using real morphed images.

corresponding to specific facial regions, concentrating the simulated artifacts in areas like the eyes and mouth where they are most common. This process yields a "morphed source" image (I_{MS}) containing localized, pixel-level irregularities.

Finally, to simulate the subtle fingerprints left by advanced generative models, the pipeline introduces frequency-based artifacts. This is achieved by generating a random, structured pattern (such as a grid or stripes), converting it to the frequency domain using a Fast Fourier Transform (FFT), and superimposing its frequency magnitudes onto the spectrum of the I_{MS} image. An inverse FFT then transforms the modified spectrum back into a final image (I_{FMS}), which now contains abnormal frequency patterns not present in pristine images. The manipulated images from these last two stages (I_{MS} and I_{FMS}) constitute the "attack" class, which, when trained against the bona fide class, teaches the model to recognize a broad and generalized set of manipulation features.

The final training dataset for the classifier is composed of the original and augmented bona fide images ($I_{\mathrm{OS}}, I_{\mathrm{AS}}$) as the authentic class, and the pixel- and frequency-manipulated images ($I_{\mathrm{MS}}, I_{\mathrm{FMS}}$) as the attack class, maintaining a balanced 1:1 ratio. This self-supervised approach forces the model to learn a rich, generalized understanding of what constitutes a manipulation artifact, rather than memorizing the quirks of a specific dataset.

3.4.2 Model Fine-Tuning with Low-Rank Adaptation (LoRA)

While full fine-tuning of large foundation models like *Gemma-3* offers a direct path to adapting them for specific downstream tasks, it is an exceptionally resource-intensive process. It requires updating all model parameters (in this case, 12 billion), which not only demands significant computational power and memory but also results in a new, large set of model weights for each task. This approach is often infeasible and inefficient. Furthermore, aggressive updates to the entire model can sometimes lead to "catastrophic forgetting," where the model loses some of the powerful, generalized knowledge acquired during its initial pre-training [59].

To circumvent these challenges, we employ a parameter-efficient fine-tuning (PEFT) strategy known as Low-Rank Adaptation (LoRA) [16]. LoRA provides an effective and computationally efficient method for adapting large pre-trained models by drastically reducing the number of trainable parameters, without introducing any inference latency.

The LoRA Mechanism. The core hypothesis underpinning the LoRA methodology is that the change in a model's weights during adaptation to a new task has a low "intrinsic rank". This suggests that the weight update matrix, which represents the difference between the initial pre-trained weights and the final fine-tuned weights, can be effectively approximated by a low-rank decomposition. Instead of updating the dense, high-dimensional weight matrix directly, LoRA freezes the original pre-trained weights and injects a pair of small, trainable "rank decomposition" matrices alongside them.

Formally, for a given pre-trained weight matrix W_0 of dimensions $d \times k$, its update ΔW is represented by two smaller matrices, A (with dimensions $r \times k$) and B (with dimensions $d \times r$), where the rank r is significantly smaller than d and k ($r \ll \min(d, k)$). During the forward pass, the output of the adapted layer is modified by combining the output of the original frozen weights with the output of these new low-rank matrices. The modified forward pass can be expressed as

$$h = W_0 x + \Delta W x = W_0 x + BAx, \tag{3.1}$$

as shown in Equation (3.1). Here, W_0 is the original pre-trained weight matrix,

which remains frozen and does not receive gradient updates during training. The matrices A and B contain the only trainable parameters for this layer. This reparameterization is highly efficient; for a large matrix W_0, the number of trainable parameters is reduced from $d \times k$ to $r \times (d + k)$. For our implementation with $r = 16$, this results in a reduction of trainable parameters by several orders of magnitude. At the start of training, A is typically initialized with a random Gaussian distribution, and B is initialized to zero. This ensures that the initial update

$$\Delta W = BA, \tag{3.2}$$

as shown in Equation (3.2), is zero, so the model's state at the beginning of the adaptation process is identical to its pre-trained state.

A scaling factor, α, is often applied to the low-rank update. In our work, we follow this practice, and the update is scaled by $\frac{\alpha}{r}$. This decouples the learning rate from the choice of rank and helps stabilize training by controlling the magnitude of the adaptation. After training is complete, the weights can be merged for deployment by explicitly calculating

$$W = W_0 + \frac{\alpha}{r} \cdot B \cdot A, \tag{3.3}$$

as shown in Equation (3.3), ensuring that no additional parameters or computational steps are added during inference.

LoRA Implementation for Morphing Attack Detection. In our approach, we applied LoRA adapters to the **Gemma-3 12B model**. To ensure a comprehensive adaptation for the morphing attack detection task, which involves both visual feature extraction and contextual understanding, LoRA matrices were injected into key components of the model's architecture. Specifically, we adapted the **query (q) and value (v) projection matrices** in every self-attention layer of both the vision tower and the entire language model. This follows best practices from prior work [4], which has shown that adapting the attention mechanism is a highly effective way to steer a model's focus toward task-specific features. The LoRA hyperparameters were set to a **rank of** $r = 16$ and a **scaling factor of** $\alpha = 32$. All LoRA parameters were trained using the AdamW optimizer with standard $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1\text{e}{-8}$.

Our approach involves augmenting the *Gemma-3* architecture with a dedicated classification head. This head is a simple, fully-connected neural network layer appended to the final output of the model, which is trained to produce a two-dimensional output corresponding to the "bona fide" and "morph" classes. During training, the original weights of the *Gemma-3* model were kept frozen, while the LoRA adapter matrices and the new classification head were trained concurrently. The entire framework was optimized using the **Binary Cross-Entropy (BCE) loss**, a standard loss function for binary classification tasks. The BCE loss is calculated as

$$L_{\text{BCE}} = -\left[y \cdot \log(p) + (1-y) \cdot \log(1-p)\right], \tag{3.4}$$

where y is the ground truth label (0 for bona fide, 1 for morph), and p is the probability of the sample being a morph, as predicted by the classification head. To facilitate stable and effective training, we employed a differential learning rate scheme. A very low learning rate of 6e−6 was used for the vision tower and 3e−7 for the language model, ensuring that the powerful, pre-trained backbone was updated delicately. A much higher learning rate of 1e−4 was used for the small, randomly-initialized classification head, allowing it to learn the classification task rapidly from scratch.

In summary, the adaptation of the *Gemma-3* model was accomplished using the parameter-efficient LoRA framework. This strategy was chosen to effectively retarget the model's pre-trained knowledge towards the specialized task of morphing attack detection while avoiding the prohibitive costs and risks of full fine-tuning. The application of LoRA was comprehensive, targeting the self-attention mechanisms across both the vision and language components. This methodology allows for a comprehensive analysis of how a large foundation model can be best adapted for this specific security application. The empirical results of this adaptation, detailing the performance on the designated evaluation datasets, are presented in the subsequent chapter.

4 Experiments

This chapter presents a comprehensive empirical evaluation of multimodal large language models for face morphing attack detection, encompassing zero-shot performance assessment, prompt engineering optimization, LoRA fine-tuning, and comparative analysis against established baseline methods. The evaluation examines four state-of-the-art multimodal LLMs across eight diverse datasets using standardized ISO/IEC 30107-3 metrics [60] to assess detection capabilities and practical deployment considerations.

4.1 Datasets

A number of specialized datasets have been created to train and evaluate MAD algorithms, each with different morph generation techniques and evaluation protocols. We summarize the most prominent public datasets below, including their creation methods and relevance.

4.1.1 Evaluation Dataset Overview

Our comprehensive evaluation employs eight specialized datasets representing diverse morphing generation techniques and evaluation protocols. These datasets are organized into two categories based on their usage: primary evaluation datasets used exclusively for testing model performance, and dual-purpose datasets that serve both evaluation and baseline training according to the protocol used by SPL-MAD [6].

4.1.1.1 Primary Evaluation Datasets

The following datasets were used exclusively for evaluation of our proposed methods and serve as the primary benchmarks for assessing morphing attack detection performance across different morphing techniques and imaging conditions.

FRGC-Morphs: The Face Recognition Grand Challenge Morphs dataset contains morphs generated from the FRGC v2 face dataset[23] [61], as illustrated in Figure 4.2. Researchers created this by pairing subjects with similar appearance and producing morphs using OpenCV, FaceMorpher, and StyleGAN2 for each pair. In total, FRGC-Morphs provides 964 morphed images covering those three techniques. The dataset's structure is organized for vulnerability analysis of face recognizers: it includes scripts to integrate with the original FRGC data and defines protocols for using morphs as either gallery or probe images in verification tests. For MAD research, FRGC-Morphs serves as a standard evaluation set, especially to assess performance on GAN-based vs classical morphs. It was used in studies like Sarkar et al. (2020) [61], which examined how GAN-morphs (MorGAN) compare to landmark morphs in fooling FR systems. Although distribution of FRGC-Morphs has been suspended (per Idiap), it remains referenced in literature and the statistics (number of morphs and types) are often reported [61]. Together with FERET-Morphs and FRLL-Morphs, it forms a triad of benchmark datasets each with around one thousand morphed images representing multiple generation algorithms.

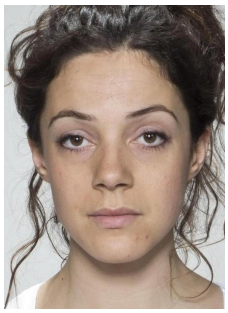
FRLL-Morphs: The Face Research Lab London Morphs dataset was derived from high-quality frontal face images of the FRL London database [22][61][62], as shown in Figure 4.2. It contains morphs created by five distinct techniques: StyleGAN2 morphs (GAN-based blending)[52], WebMorph (landmark-based web tool) [63], AMSL (an open-source landmark-based method from Hochschule Darmstadt) [64], FaceMorpher (another automated morphing software), and OpenCV (classical image warping using facial landmarks), three of which are illustrated in Figure 4.1. Each technique subset includes 1,222 morphed images and a smaller set of 204 bona fide images. Notably, FRLL-Morphs does not provide a predefined train-test split – it is generally used for evaluation only (to test generalization)

(a) OpenCV morph (b) FaceMorpher morph (c) WebMorph morph

Figure 4.1: Landmark-based morphing techniques from FRLL dataset: (a) OpenCV morphs using Delaunay triangulation with visible geometric artifacts, (b) FaceMorpher morphs with automated blending algorithms, and (c) WebMorph morphs created through web-based landmark detection, all exhibiting characteristic pixel-level artifacts such as ghosting and texture discontinuities.

because all images come from the same source set and the total number of identities is limited. FRLL-Morphs is a crucial benchmark because it offers a wide variety of morphing approaches under consistent image conditions (studio-quality images, controlled pose), enabling researchers to assess how detectors perform on different morph types[11].

MIPGAN-II. The *MIPGAN-II* dataset was created by *Zhang et al.* using face images from the FRGC-V2 database, containing morphs generated from 140 unique subjects (47 female, 93 male) with each subject contributing 7–21 additional samples for a total of 1,270 bona fide images [23, 26]. The MIPGAN-II technique generates high-resolution (1024×1024) morphed images, as shown in Figure 4.3, using StyleGAN2 architecture with a customized loss function incorporating perceptual quality, identity factors, and structural similarity measures [52, 26]. The approach differs from MIPGAN-I by utilizing the improved StyleGAN2 foundation, resulting in enhanced image quality and reduced artifacts [52, 26]. The dataset includes 1,751 MIPGAN-II morphed images alongside the corresponding bona fide samples, evaluated across three conditions: digital images, print-scanned images (using DNP-DS820 dye-sublimation printer and Canon office scanner at 300 dpi), and print-scanned images compressed to 15kb [26]. The morphing process employs identity-prior driven optimization with



(a) FRLL

(b) FRGC

(c) FERET

Figure 4.2: Dataset diversity across bona fide sources: (a) FRLL provides high-quality frontal studio conditions, (b) FRGC offers more varied backgrounds and illumination, and (c) FERET represents legacy image characteristics as well as more diverse demographics, enabling comprehensive cross-dataset generalization assessment.

weighted linear averaging of latent vectors from contributing subjects, followed by synthesis network generation and multi-component loss function optimization [26]. This dataset has demonstrated high attack success rates against both commercial and deep learning-based face recognition systems across different operational scenarios [26].

FERET-Morphs: This dataset is built from the older FERET face collection[24], as shown in Figure 4.2, by selecting look-alike pairs and generating three types of morphs per pair [61][62]. The morphing tools used are the same trio as in FRGC-Morphs (OpenCV, FaceMorpher, StyleGAN2). *Sarkar et al. (2022)* introduced FERET-Morphs to complement FRGC-Morphs in evaluating vulnerability to GAN-based vs landmark-based morphs[62]. FERET-Morphs contains 529 morphed images (for each technique), somewhat smaller in scale than FRGC, owing to the number of suitable pairs available in FERET [51]. Like FRGC/FRLL, it is usually used in conjunction with its source dataset for analysis and does not have a dedicated train/test split. The importance of FERET-Morphs lies in its inclusion of older image data and demographics, helping test detectors on different image qualities and distribution than FRLL or FRGC [65][51].

MorDIFF. *MorDIFF* was introduced by *Damer et al.* (2023) to investigate the potential of diffusion autoencoders for creating representation-level face morphing attacks, addressing the limitations of GAN-based approaches in reconstruction fidelity and identity preservation [29]. Built upon the *Face Research Lab London (FRLL)* dataset [22], *MorDIFF* extends the *SYN-MAD 2022* competition dataset by using identical morphing pairs to enable direct comparison with existing techniques [3].

The dataset employs a novel diffusion-based morphing approach that interpolates both semantic and stochastic latent representations: linear interpolation (Lerp) for semantic features and spherical linear interpolation (SLerp) for stochastic components [66, 67], before decoding the combined latent code through a conditional diffusion probabilistic model [68, 69]. *MorDIFF* contains 1,000 morphing attack images generated from 1,000 carefully selected pairs (250 each for female neutral, female smiling, male neutral, and male smiling expressions) alongside 204 bona fide images from *FRLL*. Pair selection utilized *ElasticFace-Arc* embeddings [70] with cosine similarity matching within gender and expression splits to identify the most similar faces.

Vulnerability analysis demonstrated that *MorDIFF* attacks achieve significantly higher Mated Morph Presentation Match Rates (MMPMR) than existing GAN-based representation-level morphs (MIPGAN I/II [26]), with performance approaching that of traditional image-level morphing techniques while exhibiting superior visual quality and reduced generation artifacts, as demonstrated in Figure 4.3. The dataset represents a critical advancement in understanding next-generation morphing threats, as diffusion-based attacks combine the identity preservation strength of image-level morphs with the artifact reduction benefits of representation-level techniques.

Greedy-DiM. *Greedy-DiM* morphs were created by *Blasingame et al.* using diffusion-based morphing with greedy optimization strategies [25]. Built from the *SYN-MAD 2022* competition dataset [3], which derives from the *Face Research Lab London (FRLL)* dataset containing 102 individuals with high-quality frontal images under neutral lighting conditions [22], the dataset employs 489 bona fide image pairs (reduced from 500 due to technical issues). Pair selection was con-

ducted using the *ElasticFace* face recognition system [70], identifying the top 250 most similar pairs for each gender based on cosine similarity.

The dataset includes morphs generated using two variants: *Greedy-DiM-S* (greedy search strategy) and *Greedy-DiM** (greedy optimization strategy), both leveraging diffusion models with iterative sampling processes guided by identity-based heuristic functions [25]. In this study, we utilized the *Greedy-DiM* variant for evaluation, as illustrated in Figure 4.3.*Greedy-DiM** achieves optimization through gradient descent on noise predictions at each timestep, requiring only 270 Network Function Evaluations (NFEs) compared to 2,350 for competing methods like *Morph-PIPE* [30].

The approach demonstrated unprecedented attack effectiveness, achieving 100% Mated Morph Presentation Match Rate (MMPMR) across *ArcFace* [71], *AdaFace* [72], and *ElasticFace* recognition systems. Images are processed at 256×256 resolution with alignment and cropping following the preprocessing pipeline of the *FFHQ* dataset [73], making this dataset particularly significant for evaluating next-generation diffusion-based morphing attacks that consistently outperform both landmark-based and GAN-based methods.

(a) MIPGAN-II (b) Greedy-DiM (c) MorDIFF

Figure 4.3: Advanced morphing techniques in evaluation datasets: (a) MIPGAN-II represents sophisticated GAN-based synthesis with identity-aware optimization, (b) Greedy-DiM demonstrates cutting-edge diffusion-based morphing with exceptional seamlessness, and (c) MorDIFF illustrates diffusion autoencoder interpolation approaches.

4.1.1.2 Dual-Purpose Datasets

These datasets served a dual role in our evaluation framework: as test sets for performance assessment and as training data for supervised baseline implementations (PW-MAD [5], MixFaceNet-MAD [21], Inception-MAD [5]) following the cross-dataset evaluation protocol established in SPL-MAD [6].

MorGAN. *MorGAN* was introduced by *Damer et al.* (2018) as the first comprehensive evaluation framework for GAN-based representation-level face morphing attacks, representing a paradigm shift from traditional landmark-based approaches to deep learning-generated morphs [27]. Built upon the *CelebA* dataset [74] with extensive filtering to ensure ICAO-compliant frontal face images [75], *MorGAN* employs a novel generative adversarial network (GAN) architecture that combines encoder–decoder components with adversarial training to achieve superior identity preservation.

The dataset contains 1,000 *MorGAN* attacks, 1,000 landmark-based morphing (LMA) attacks for comparison, 1,500 bona fide reference images (500 key references and 1,000 secondary references), and 1,500 corresponding probe images, totaling 5,000 images across identity-disjoint train and test splits [27], as exemplified in Figure 4.4. The *MorGAN* approach performs morphing in latent space by encoding source images, linearly interpolating their representations with $\beta = 0.5$, and decoding the result through a generator enhanced with pixel-wise reconstruction loss ($\alpha = 0.3$) to preserve facial identity information [27].

Despite technical limitations requiring 64×64 pixel resolution, vulnerability analysis demonstrated that *MorGAN* attacks successfully compromise face recognition systems while exhibiting different detectability characteristics compared to traditional morphing methods [27]. The dataset's significance lies in pioneering representation-level morphing attacks and establishing the foundation for GAN-based face morphing research, highlighting the evolving threat landscape that detection systems must address.

(a) MorGAN-GAN (b) MorGAN-LMA (c) Bona fide

Figure 4.4: MorGAN dataset comparison at original resolution (64×64 pixels): (a) GAN-based morph created through encoder-decoder architecture with latent space interpolation, (b) landmark-based morph (LMA) using traditional geometric warping for direct comparison, and (c) genuine bona fide image.

LMA-DRD. *LMA-DRD* (Digital and Re-digitized Landmark-based Morph Dataset) was created by *Damer et al.* (2021) to address the critical gap in evaluating morphing attack detection performance on re-digitized images, which reflects real-world passport issuance scenarios where printed photos are scanned [5]. Built from the *VGGFace2* dataset [76], *LMA-DRD* contains carefully filtered frontal face images meeting International Civil Aviation Organisation (ICAO) travel document requirements [77]. The dataset employs landmark-based morphing techniques following the pipeline described in [78], pairing 197 key images with their most similar counterparts based on *OpenFace* similarity measurements [79].

LMA-DRD provides 276 digital morphing attacks (D-M) and 364 digital bona fide images (D-BF), with corresponding re-digitized versions (PS-M and PS-BF) created through professional printing on glossy photo paper and scanning at 600 dpi, as shown in Figure 4.5. The dataset is organized into identity-disjoint train, development, and test splits to prevent evaluation bias. Vulnerability analysis using *ResNet-100 ArcFace* [71] demonstrates high Mated Morph Presentation Match Rates (MMPMR) of 91.30% for digital attacks and 88.41% for re-digitized attacks at 1.0% FMR (as recommended for border check operations [80]) [5].

LMA-DRD's unique contribution lies in its systematic investigation of the print-scan degradation effect on morphing attack detection, making it essential for evaluating detector generalization to real-world deployment scenarios where physical document processing introduces additional artifacts and challenges.

(a) LMA-DRD Digital (b) LMA-DRD Print-Scan

Figure 4.5: LMA-DRD dataset examples: (a) Digital morphing attack from the LMA-DRD Digital subset, and (b) corresponding print-scan version from the LMA-DRD Print-Scan subset, demonstrating how the physical printing and scanning process affects the image.

4.1.2 Training Dataset Overview

The SMDD dataset served distinct training roles for different model categories in our evaluation framework, requiring different data utilization strategies based on the underlying training methodology.

4.1.2.1 Supervised Baseline Training

For supervised baseline implementations (PW-MAD [5], MixFaceNet-MAD [21], Inception-MAD [5]), the complete SMDD dataset was utilized following conventional supervised training protocols, employing both original morphed images and bona fide samples with explicit binary labels.

SMDD. *SMDD* (Synthetic Morphing Attack Detection Development) was introduced by *Damer et al.* (2022) as the first synthetic-based biometric attack detection dataset, addressing critical legal and privacy challenges associated with using real biometric data for MAD development under GDPR regulations [28]. Generated using *StyleGAN2-ADA* [81] trained on the *Flickr-Faces-HQ (FFHQ)* dataset [73], *SMDD* begins with 500,000 randomly generated synthetic face images from Gaussian noise vectors, which are subsequently filtered to the highest-

quality 50,000 samples using *CR-FIQA* quality assessment [82] to remove extreme poses and occlusions.

The dataset employs a systematic construction methodology where 25,000 images serve as bona fide samples, while 5,000 key morphing images are each paired with five randomly selected images from the remaining 20,000 samples, creating 25,000 morphing pairs to maximize training diversity rather than attack strength [28]. Morphing attacks are generated using the widely adopted *OpenCV*/dlib landmark-based algorithm with Delaunay triangulation [83, 84], followed by manual quality filtering to remove artifacts, resulting in 15,000 high-quality morphing attacks. The complete dataset provides 40,000 training samples (25,000 bona fide + 15,000 attacks) and 40,000 evaluation samples with identical composition, totaling 80,000 images across identity-disjoint splits.

SMDD's groundbreaking contribution lies in demonstrating that privacy-friendly synthetic data can successfully train MAD systems that generalize effectively to unknown real-world morphing techniques, potentially revolutionizing biometric security research by eliminating the legal, ethical, and scalability constraints of real biometric data usage.

4.1.2.2 LoRA Fine-Tuning Data

Our Gemma-3 12B LoRA adaptation employed exclusively the bona fide subset of SMDD (25,000 images), implementing a self-supervised framework inspired by SelfMAD [1] to synthetically generate morphing artifacts during training, eliminating dependence on pre-existing morphed samples.

Self-Supervised Artifact Generation Following the SelfMAD methodology [1], the training process simulates morphing artifacts through a three-stage pipeline: (i) image augmentation applying color, brightness, and quality variations; (ii) pixel-artifact generation using geometric transformations and blending to replicate landmarks-based morphing irregularities; and (iii) frequency-artifact generation superimposing structured patterns in the frequency domain to simulate GAN and diffusion-based morphing fingerprints.

This approach enables the model to learn generalizable morphing detection features without overfitting to specific attack techniques, as demonstrated by SelfMAD's superior cross-dataset generalization compared to supervised approaches [1]. The synthetic artifact generation produces both pixel-space and frequency-space irregularities, allowing classification of genuine images (original and augmented) versus manipulated samples (pixel and frequency artifacts) with a balanced 1:1 ratio.

4.1.3 Image Preprocessing

Our experimental framework required distinct preprocessing approaches for different evaluation scenarios, reflecting the varied requirements of baseline comparison methods, zero-shot multimodal LLM evaluation, and fine-tuning procedures. This section details the three preprocessing pipelines employed to ensure appropriate data preparation for each experimental condition.

Baseline Models Preprocessing. For benchmark training and evaluation of classical MAD methods, we adhered to the preprocessing specifications established in the original publications to ensure fair comparison and reproducibility. Each baseline method utilized its documented preprocessing requirements, including specific image resolution standards, normalization procedures, and augmentation strategies as reported in their respective papers. Supervised methods such as **MixFaceNet-MAD**, **Inception-MAD**, and **PW-MAD** were trained using their published preprocessing pipelines [21, 5], with images resized to model-specific dimensions and normalized according to their training protocols. Unsupervised approaches including **SPL-MAD** and **MAD-DDPM** employed their original preprocessing configurations [6, 40] to maintain consistency with reported performance metrics.

Zero-Shot Multimodal LLM Preprocessing. For zero-shot evaluation of multimodal large language models, we provided raw images to leverage each model's native preprocessing capabilities, allowing their built-in processors to handle format conversion and optimization. **Gemma-3 27B Vision** employs a

SigLIP vision encoder operating on fixed 896×896 square images, utilizing a "Pan&Scan" algorithm to handle different aspect ratios and high resolutions by adaptively cropping and resizing images [54, 85]. **Qwen2.5-VL 32B** supports dynamic resolution inputs with configurable `min_pixels` and `max_pixels` parameters (default: `min_pixels` $= 256 \times 28 \times 28$, `max_pixels` $= 1280 \times 28 \times 28$), with dimensions rounded to the nearest multiple of 28 [18]. **Llama-4-Scout 17B** [19] incorporates early fusion for native multimodality through its `AutoProcessor`, handling various input formats including image URLs and base64-encoded images [19]. **Mistral Small 3.1 24B** utilizes the `mistral_common` library with vLLM processing, supporting multimodal input through `ImageURLChunk` and `TextChunk` components while requiring base64 format for certain API configurations [56, 86]. This approach ensured that each model operated under its optimal preprocessing conditions without introducing artifacts from manual preprocessing steps.

Gemma-3 Fine-Tuning Preprocessing. For LoRA fine-tuning experiments with **Gemma-3** [54], we implemented a sophisticated preprocessing pipeline specifically designed for morphing attack detection training. Images were processed at 896×896 pixel resolution to match the **SigLIP** vision encoder requirements [85]. The pipeline incorporated a multi-stage synthetic attack generation system that created both genuine and manipulated training examples from the **SMDD** dataset [28]. Self-blending techniques applied geometric transformations, color space modifications, and statistical averaging effects to simulate morphing artifacts. Frequency-domain enhancements introduced spectral artifacts through Fourier transform manipulation, while extensive data augmentation included horizontal flipping, random cropping with variable margins, color space perturbations, and JPEG compression simulation. Images underwent normalization using SigLIP-compatible parameters, with pixel values scaled to the $[0, 1]$ range and subsequently normalized with means and standard deviations of [0.5, 0.5, 0.5]. The preprocessing concluded with tokenization using structured prompts that established forensic analysis context, with sequences padded to accommodate the model's 2048 token limit.

4.1.4 Dataset Partitioning

To maintain experimental integrity, we implemented a carefully structured dataset partitioning approach that accommodates multiple training and evaluation scenarios. For supervised baseline methods, we followed their original training protocols as established in the literature. Specifically, *MixFaceNet-MAD*, *PW-MAD*, and *Inception-MAD* baselines were trained on five distinct datasets following the evaluation framework of Fang et al. [6]: *SMDD* [28], *MorGAN-LMA* [27], *MorGAN-GAN* [27], *LMA-DRD (Digital)* [5], and *LMA-DRD (Print-Scan)* [5], with each architecture trained independently on each dataset to assess cross-dataset generalization capabilities. For our multimodal LLM fine-tuning experiments, the *bona fide* subset of the *SMDD* dataset (25,000 images) [28] was used with the *SelfMAD* self-supervised training methodology [1] for LoRA adaptation of *Gemma-3* [54], ensuring no overlap with zero-shot evaluation data. The remaining datasets, *FRLL-Morphs* [22], *FRGC-Morphs* [23], *FERET-Morphs* [24], *Greedy-DiM* [25], *MIPGAN-II* [26], and *MorDiff* [29], served as evaluation benchmarks for both zero-shot multimodal LLM assessment and comprehensive baseline comparison. This partitioning strategy prevented data leakage between training and evaluation phases while enabling rigorous cross-dataset generalization analysis, providing reliable assessment of model performance across diverse morphing techniques and generation methods that were unseen during training.

4.2 Evaluation Metrics

To rigorously assess the performance of our morph detection models, we adopt the standardized metrics defined by the ISO/IEC 30107-3 standard for biometric presentation attack detection [60]. These metrics are designed to quantify a model's ability to correctly classify both bona fide (genuine) and attack (morphed) images. Given that our model produces a continuous suspicion score, its performance is evaluated across a range of decision thresholds (τ).

The two primary error rates are the *Bona Fide Presentation Classification Error Rate (BPCER)* and the *Attack Presentation Classification Error Rate (APCER)*.

Bona Fide Presentation Classification Error Rate (BPCER). BPCER measures the proportion of bona fide presentations that are incorrectly classified as morphing attacks. This is analogous to the False Positive Rate (FPR) in a standard binary classification context. It is calculated as

$$\mathrm{BPCER}(\tau) = \frac{N_{\mathrm{BF}\to\mathrm{A}}}{N_{\mathrm{BF}}}, \tag{4.1}$$

where $N_{\mathrm{BF}\to\mathrm{A}}$ is the number of bona fide images with a morph suspicion score exceeding the threshold τ, and N_{BF} is the total number of bona fide images.

Attack Presentation Classification Error Rate (APCER). APCER measures the proportion of morphing attack presentations that are incorrectly classified as bona fide. This corresponds to the False Negative Rate (FNR). It is calculated as

$$\mathrm{APCER}(\tau) = \frac{N_{\mathrm{A}\to\mathrm{BF}}}{N_{\mathrm{A}}}, \tag{4.2}$$

where $N_{\mathrm{A}\to\mathrm{BF}}$ is the number of attack images with a morph suspicion score below the threshold τ, and N_{A} is the total number of attack images.

We conducted detailed analysis at fixed operating points to assess model performance across various security thresholds. Specifically, we measured APCER at fixed BPCER thresholds of 0.01%, 1%, 5%, 10%, and 20%, as well as BPCER at corresponding fixed APCER thresholds. This multi-point analysis enables practical deployment considerations where different applications may require varying false positive and false negative tolerance levels.

Equal Error Rate (EER). There is an inherent trade-off between APCER and BPCER, controlled by the decision threshold. To provide a single, threshold-independent performance summary, the *Equal Error Rate (EER)* is commonly used. EER is the point at which the APCER and BPCER are equal. A lower EER value indicates superior overall performance, as it represents a better balance between correctly identifying attacks and not misclassifying genuine images. The EER is found at the threshold τ^* where

$$\mathrm{APCER}(\tau^*) = \mathrm{BPCER}(\tau^*). \tag{4.3}$$

Finally, to provide the most comprehensive assessment of a model's discriminative power across all possible thresholds, we use the *Area Under the Receiver Operating Characteristic Curve (AUC-ROC)* as our evaluation metric. The ROC curve is generated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR).

Area Under the Receiver Operating Characteristic Curve (AUC-ROC). The AUC-ROC provides a comprehensive, threshold-independent measure of the model's ability to distinguish between bona fide and morphed images. The True Positive Rate (TPR) is defined as

$$\text{TPR} = 1 - \text{APCER}, \tag{4.4}$$

while the False Positive Rate (FPR) is equivalent to the Bona Fide Presentation Classification Error Rate (BPCER):

$$\text{FPR} = \text{BPCER}. \tag{4.5}$$

The area under this curve, or AUC, represents the probability that the model will assign a higher suspicion score to a randomly chosen morphing attack image than to a randomly chosen bona fide image. An AUC of 1.0 indicates a perfect classifier, able to flawlessly separate attack and bona fide presentations, whereas an AUC of 0.5 suggests performance no better than random chance. The AUC is calculated by integrating the ROC curve:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d\text{FPR}. \tag{4.6}$$

A key advantage of the AUC-ROC metric in morphing attack detection is that it summarizes the model's discriminative power over its entire operating range, unlike the Equal Error Rate (EER) which evaluates performance at a single operating point.

The comprehensive evaluation protocol ensures robust assessment of morphing attack detection capabilities while providing detailed insights into model behavior across diverse operational requirements and facilitating meaningful comparison with established state-of-the-art methods.

4.3 Zero-Shot Experimental Setup

To ensure a robust and replicable evaluation, we established a precise inference configuration for all models. A key objective was to balance deterministic output for consistency with the analytical flexibility required for the nuanced task of morph detection.

We configured each model with a *temperature* of 0.1. This value was chosen through experimentation; a *temperature* of 0.0, corresponding to pure greedy decoding, was found to be overly restrictive, limiting the models' ability to generate the detailed, reasoned analysis required by our more complex prompts. Conversely, a *temperature* of 0.1 provided a high degree of consistency, ensuring that the core analytical outcomes (e.g., final suspicion scores, identified artifacts, and overall judgment) remained stable across multiple iterations of the same query. While this setting could result in trivial variations in the phrasing of the natural language rationale fields, the critical quantitative and qualitative findings were reliably reproducible. The *top_p* parameter was not modified and was left at its default value. To manage output length, the *max_new_tokens* parameter was adjusted based on the prompt type: a small limit (e.g., 1–5 tokens) was used for prompts expecting a single-word or score-only answer, while a much larger limit was set for the structured JSON output to ensure the full, unabridged response could be generated.

The evaluation process was conducted sequentially for each model to ensure controlled and isolated measurement. For a given model hosted on the 4-GPU server, each image from the test datasets (described in Section 4.1.1) was processed individually against the chosen prompt variant. There were no parallel queries processing different images or prompts simultaneously. The model's complete, raw text output was meticulously recorded for every (image, prompt) pair, forming the basis for our results analysis. Given the sequential nature of this workflow, the total runtime to evaluate a single model across the entire dataset was in the order of a couple of days up to a week. The inference times varied significantly across the four multimodal LLMs due to their different architectural complexities and optimization strategies. As shown in Table 4.1 Gemma-3 27B required approximately 30 seconds per image, Qwen2.5-VL 32B required 25 sec-

onds per image, Mistral Small 3.1 24B demonstrated the fastest processing time at 9 seconds per image, while Llama-4-Scout 17B required 40 seconds per image for complete inference including prompt processing and response generation.

Table 4.1: Average inference times for different multimodal LLMs in zero-shot evaluation setting.

Model	Parameters	Inference Time (seconds/image)
Gemma-3 27B	27B	30
Qwen2.5-VL 32B	32B	25
Llama-4-Scout 17B	17B	40
Mistral Small 3.1 24B	24B	9

4.4 LoRA Fine-Tuning Setup

Following the comprehensive zero-shot evaluation of multimodal large language models, we conducted systematic fine-tuning experiments to investigate the potential performance gains achievable through domain-specific adaptation. These experiments focused on adapting the *Gemma-3 12B* model using Low-Rank Adaptation (LoRA) techniques, leveraging the self-supervised training methodology established in the SelfMAD framework to create synthetic morphing attack detection training data.

For our experiment, we augmented the pre-trained *Gemma-3* architecture with a dedicated binary classification head. This adaptation allowed the model to be optimized for binary classification through the use of a **Binary Cross-Entropy (BCE)** loss function, as described in Equation (3.4).

4.4.1 Training and Validation Setup

Two separate training runs were conducted with the augmented *Gemma-3* architecture. The first run established a baseline by training only the classification head for 2 epochs, with no *LoRA* parameters. This baseline training used the

same setup as the main experiment: a learning rate of 1×10^{-4} for the classification head, a batch size of 2 with gradient accumulation of 16, and a 500-step warm-up. The second run then involved comprehensive fine-tuning with *LoRA* adapters added to the entire model.

The Low-Rank Adaptation (*LoRA*) fine-tuning was executed on a high-performance computing platform equipped with two *NVIDIA A100 80GB* GPUs, where the 12.2-billion parameter *Gemma-3* model was distributed using model parallelism to ensure efficient training. *LoRA* adapters were applied to both the vision and language components of the model, resulting in 74,657,537 trainable parameters, which constitutes 0.61% of the model's total parameters.

A differential learning rate strategy was implemented to fine-tune the model's components appropriately. The newly added classification head was trained with a learning rate of 1×10^{-4}, while the language and vision *LoRA* parameters were updated more conservatively with learning rates of 7×10^{-6} and 9×10^{-6}, respectively. This approach allows for the rapid adaptation of the task-specific classification layer while preserving the robust features of the pre-trained base model.

Training was configured with a batch size of 2 and gradient accumulation of 16, yielding an effective global batch size of 32. The training schedule was set for 30 epochs and included a 500-step warm-up phase followed by a linear learning rate decay.

Validation Performance Monitoring. To track generalization performance and prevent overfitting, the model was evaluated on a validation set of 2,498 samples every 400 global steps. Performance was assessed using the **Area Under the Curve (AUC)** metric, as described in Equation (4.6).

The model exhibited a rapid improvement in validation metrics within the first epoch. The AUC score increased from 0.705 at step 400 to 0.944 at step 8000. Peak performance was achieved within the second epoch, with the best model checkpoint reaching a validation AUC of 0.966 at step 25,600. The validation loss curve generally mirrored the training loss, indicating effective generalization without significant signs of overfitting.

Feature representation analysis during validation confirmed that the fine-tuning process progressively enhanced the separability of features for genuine and morphed images. The inter-class distance, a measure of the separation between the two classes in the feature space, grew from 17.4 at step 400 to a peak of 104.4 at step 24,800 (see Figure 4.6). Similarly, the separation ratio improved from 32.4 to over 68.4 in the same interval, signifying that the LoRA updates successfully refined the model's ability to produce highly discriminative representations for the morphing detection task.

This training regimen demonstrates the parameter-efficient power of LoRA in adapting a large-scale foundation model for a specialized forensic task.

Following the training phase, a comprehensive evaluation was conducted across all test datasets to quantify the performance gains achieved through domain-specific adaptation. The evaluation focused on comparing the baseline model, to the fine-tuned model incorporating *Low-Rank Adaptation (LoRA)*. This comparison highlights the impact of lightweight parameter-efficient fine-tuning technique in the context of morphing attack detection.

4.5 Classical Baselines

To contextualize the performance of our proposed method, we conducted a comprehensive evaluation against a suite of state-of-the-art and foundational benchmark models. The selected benchmarks included three prominent supervised deep learning classifiers: *MixFaceNet-MAD* [21, 28], *PW-MAD* [5], and a standard *Inception-MAD* classifier, all of which were our own re-implementations based on their original papers. The evaluation was further extended to include four recent unsupervised and self-supervised approaches. For *SPL-MAD* [6] and *MAD-DDPM* [40], we utilized the officially provided pre-trained model weights to ensure a faithful comparison to their reported results.

To ensure a rigorous analysis of generalization for the supervised models, we adopted the comprehensive evaluation framework proposed by *Fang et al.* [6]. Each of the three supervised architectures was trained independently on five distinct datasets: *LMA-DRD* (Digital and Print-Scan versions), *MorGAN* (LMA

and GAN versions), and *SMDD*. This process resulted in 15 distinct supervised model checkpoints (3 architectures $\times$ 5 training sets), enabling a thorough analysis of how training data composition affects model robustness. Finally, all benchmark models—including the 15 supervised checkpoints and the four prepared unsupervised/self-supervised models—were systematically evaluated on the full set of our evaluation datasets presented in Section4.1.1. The specific architectural details and implementation notes for each of these baseline models are provided below.

MixFaceNet-MAD. This supervised baseline is an adaptation of the *MixFaceNet* architecture, a lightweight network originally designed for face recognition that utilizes mixed depthwise convolutional blocks [21]. Following its successful application in the SYN-MAD 2022 competition [28, 3], we implemented this model to serve as a robust classification benchmark.

Inception-MAD. As a classical baseline, we included an *Inception-MAD* classifier built upon the InceptionV3 architecture [39]. For our implementation, the model was initialized with weights pre-trained on ImageNet. The final fully connected layer and the auxiliary classifier's output layer were replaced with a single neuron to perform binary classification.

PW-MAD. For this baseline, we implemented a *Pixel-Wise Morph Attack Detection (PW-MAD)* approach described in the work of *Damer et al.* [5]. Our implementation utilizes a `PW_MAD_DenseNet` architecture, which is trained to produce two outputs: a standard image-level binary classification (bona fide vs. morph) and a pixel-wise prediction map of size 14×14.

SPL-MAD. The *Self-Paced Learning MAD (SPL-MAD)* model, an unsupervised approach proposed by *Fang et al.* [6], was included as a state-of-the-art unsupervised benchmark. For our experiments, we utilized the official pre-trained model weights and code publicly released by the original authors.

This method employs a Convolutional Autoencoder (CAE) to identify morphs

as anomalies based on reconstruction error. To evaluate this model on our test sets, each image was passed through the pre-trained CAE. The decision for each image was based on its Mean Squared Error (MSE) between the input and the reconstructed output, with a higher error indicating a higher likelihood of being a morph.

MAD-DDPM. As a second state-of-the-art unsupervised baseline, we included the *MAD-DDPM* model proposed by *Ivanovska and Štruc* [40]. This approach leverages a Denoising Diffusion Probabilistic Model (DDPM) to learn the data distribution of only bona fide images and detect morphs as out-of-distribution anomalies.

For our evaluation, we used the official code and pre-trained model weights provided by the authors. According to their paper, the model was trained on the *CASIA-WebFace* dataset. We did not perform any retraining. During evaluation, each test image was processed by the pre-trained model to compute a reconstruction-based anomaly score. This score was then used to classify the image as either bona fide or a morphing attack.

4.6 Hardware Infrastructure and Computational Resources

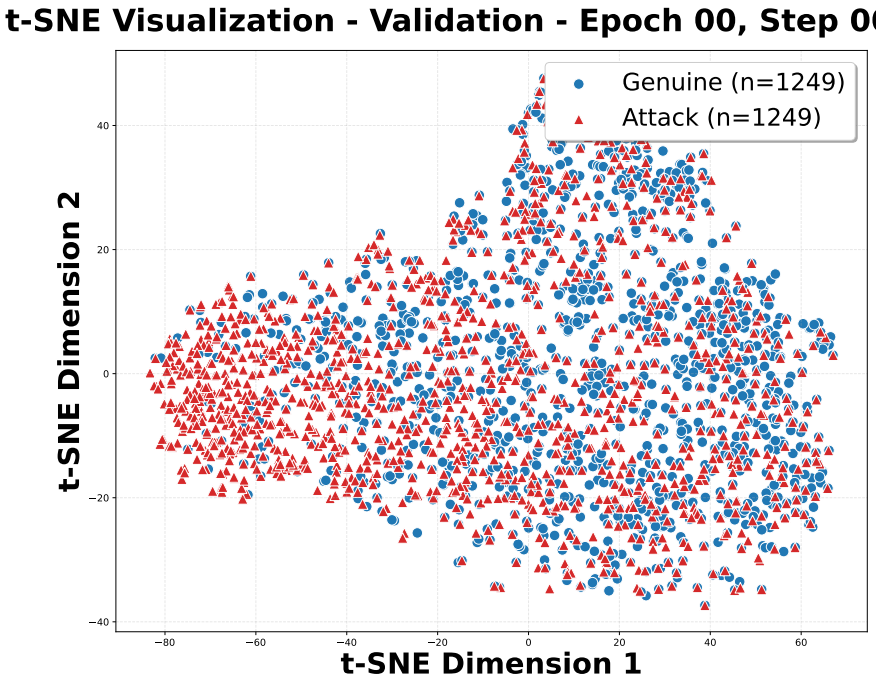
The evaluation required substantial computational resources across multiple hardware configurations. This section summarizes the practical computational requirements and performance characteristics encountered during the experimental phases.

Baseline Model Training and Evaluation. Following the configuration described in 4.5, all supervised baseline methods were trained on a single *NVIDIA RTX 4090* GPU, providing consistent computational conditions across different architectural implementations. This setup proved adequate for the memory and processing requirements of classical MAD approaches while maintaining reasonable training times.

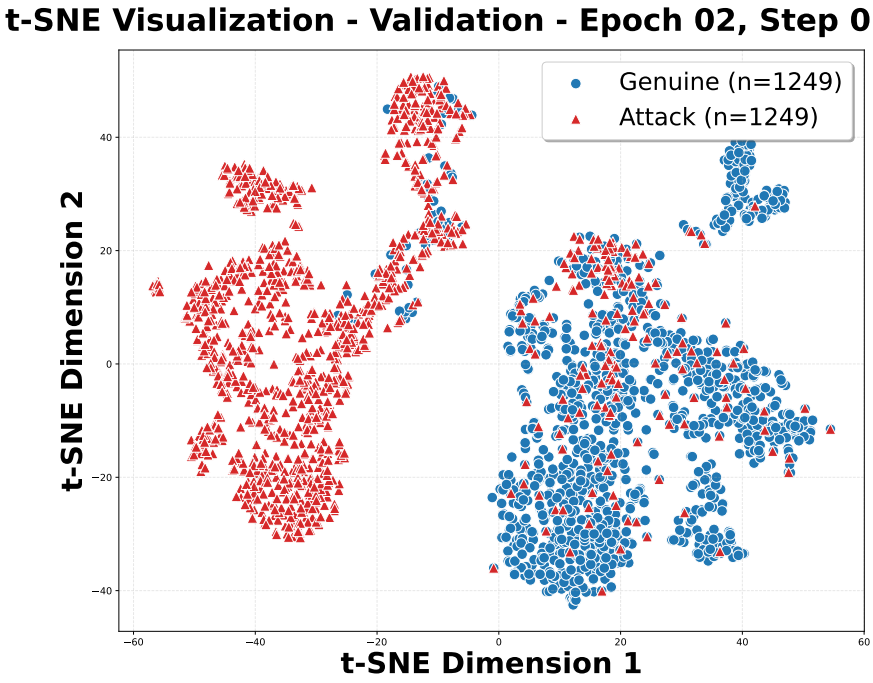
Zero-Shot Multimodal LLM Evaluation. All experiments were conducted on a uniform hardware platform to ensure comparability between models. The inference environment consisted of four NVIDIA RTX 4090 GPUs, each with 24 GB of VRAM. We utilized the vLLM inference server to host the models and efficiently manage the distributed 4-GPU setup. Due to the significant memory requirements of the M-LLMs, we employed specific precision and quantization strategies tailored to each model: *Gemma-3*, *Qwen-2.5-VL*, and *Mistral Small 3.1* were loaded and run entirely in bfloat16 precision, while the *Llama-4-Scout* model, having the largest memory footprint, was loaded using int4 quantization to fit within the available VRAM of the 4-GPU cluster. This was a necessary trade-off to enable its evaluation on our hardware.

Fine-Tuning Computational Requirements. The LoRA fine-tuning experiments, conducted on the two-GPU *NVIDIA A100 80GB* configuration, represented the most computationally intensive phase of the research. The classification head approach required 100 hours of total training time (20 hours $\times$ 5 epochs), while the generative approach demanded 300 hours (60 hours $\times$ 5 epochs). These extended training periods reflect the computational complexity of adapting large foundation models for specialized security applications.

The substantial computational investment required for this comprehensive evaluation underscores both the complexity of multimodal foundation model research and the importance of systematic experimental design to maximize the value of computational resources in advancing biometric security applications.



(a) Representation at step 400 (inter-class distance = 17.4).



(b) Representation at step 24,800 (inter-class distance = 104.4).

Figure 4.6: Visualization of inter-class separation growth in the feature space. The inter-class distance increased from 17.4 at step 400 to 104.4 at step 24,800, indicating improved class separability during training.

5 Results

This chapter presents the comprehensive evaluation results from the experimental framework described in Chapter 3. The analysis is organized in three main sections: first, we examine the zero-shot performance of four multimodal LLMs (*Gemma-3*, *Qwen2.5-VL*, *Llama-4-Scout*, and *Mistral Small 3.1*) across diverse morphing attack datasets, including the impact of prompt engineering strategies on detection accuracy (Section 5.1). Second, we present the performance improvements achieved through LoRA fine-tuning of *Gemma-3 12B*, demonstrating the effectiveness of parameter-efficient adaptation for morphing attack detection (Section 5.2). Finally, we provide a comparative analysis benchmarking our approach against established supervised, unsupervised, and foundation model baselines to contextualize the performance within the current state-of-the-art (Section 5.3). All evaluations employ the standardized ISO/IEC 30107-3 metrics [60] defined in Section 4.2, with particular emphasis on Equal Error Rate (EER) and operational performance at fixed security thresholds.

5.1 Zero-Shot Results

In this section, we evaluate the zero-shot morphing attack detection capabilities of four multimodal large language models through systematic prompt engineering optimization. We assess *Gemma-3 27B*, *Qwen2.5-VL 32B*, *Llama-4-Scout 17B*, and *Mistral Small 3.1 24B*.

5.1.1 Model Performance Across Prompt Strategies

In this section, we evaluate three prompt engineering strategies to identify the optimal approach for zero-shot morphing attack detection. This comparison establishes the prompting methodology used throughout our zero-shot experiments.

Following the iterative prompt development process described in Section 3.3, we evaluated the three distinct prompting strategies that represent critical design milestones. The complete text of all three prompts is provided in Appendices A.1–A.3.

Structured Forensic Analysis – Semantic Guide A (*Prompt 1*). Complete six-step analytical framework with simplified semantic scoring (0–10,000 scale) and basic threshold definitions. Requires structured JSON output with individual step scores and rationales.

Extended Forensic Analysis – Semantic Guide A (*Prompt 2*). Augmented version of *Prompt 3* with additional sub-questions within each analytical step and expanded contextual instructions. Provides comprehensive guidance with structured JSON format.

Optimized Forensic Analysis – Semantic Guide B (*Prompt 3*). Final refined prompt featuring detailed interpretative guide with specific score range examples, streamlined step instructions focused on artifact identification, and mandatory reasoning explanations for each component.

The performance analysis demonstrates a clear progression in prompt effectiveness across the three evaluated strategies. *Prompt 1* provided the first significant results, establishing that smaller multimodal LLMs possess inherent capability for morphing attack detection when provided with structured scoring frameworks, specific analytical steps, and appropriate forensic context. This structured approach enabled models to detect morphs with measurable accuracy, validating our hypothesis that systematic guidance could unlock latent detection capabilities.

Building on *Prompt 1's* success, *Prompt 2* attempted to enhance performance through increased structural complexity, incorporating additional sub-questions

and expanded contextual instructions within each analytical step. However, this approach yielded counterproductive results, increasing both computational overhead and average EER by 13.15% across all models. The performance degradation suggests that excessive prompt complexity can overwhelm model processing capabilities, leading to reduced rather than enhanced analytical precision.

Prompt 3 addressed these limitations through refined prompt engineering, implementing an improved semantic scoring guide and optimized six-step analytical framework that retained only the most effective instructional components. This is demonstrated in Table 5.1, this optimization resulted in an average EER reduction of 18% across all evaluation datasets, establishing *Prompt 3* as the optimal prompting strategy for morphing attack detection tasks.

Table 5.1: Detailed EER (%) comparison for *Gemma-3* with *Prompt 1* and *Prompt 3* across multiple datasets. Improvement is $\Delta\text{EER} = \textit{Prompt 1} - \textit{Prompt 3}$ (positive indicates lower error with *Prompt 3*).

Dataset	**Subset**	**P1 EER (%)**	**P3 EER (%)**	**ΔEER (% points)**
FRLL	StyleGAN2	41.4	27.4	**14.0**
	WebMorph	23.5	12.9	**10.6**
	AMSL	38.7	25.1	**13.6**
	FaceMorpher	21.7	13.1	**8.6**
	OpenCV	19.1	13.3	**5.8**
MIPGAN-II		42.9	35.6	**7.3**
Greedy_greedy_dim		18.6	6.2	**12.5**
Average		**29.4**	**19.1**	**10.3**

Notes: ΔEER is computed as P1 − P3; Positive values mean P3 improves over P1; *P1=Prompt 1*; *P3=Prompt 3*; Both prompts were employed verbatim as presented in Appendix A.1 and A.3

For all subsequent evaluations and model comparisons, *Prompt 3* was employed verbatim as presented in Appendix A.3. While initial attempts were made to adapt the prompt wording to accommodate the specific characteristics of each multimodal LLM, empirical testing revealed that the original formulation yielded superior performance across all architectures. This finding suggests that the structured analytical framework of *Prompt 3* possesses sufficient generality to ef-

fectively guide different LLMs without model-specific customization. It should be noted that *Prompt 3* was initially developed and optimized using *Qwen2.5-VL*, yet demonstrated robust transferability to all evaluated architectures.

In addition to the confidence scores, all three prompts also output the reasoning behind their decisions. The listing 5.1 provides an example of the structured JSON output, produced by *Gemma-3 27B*, for the *Prompt 3*, for the morphed image presented in Figure 5.1. The JSON structure includes a `final_decision` block containing the aggregated confidence score provided by the LLM and a qualitative summary of the findings. It also contains a `step_by_step_analysis` block, which details the score and textual rationale for each of the six forensic steps. This structure allows for a more transparent and traceable assessment of the LLMs decision-making process, linking specific visual observations to the final determination.

Figure 5.1: Example of a morphed image from the *FERET* dataset used as input for the analysis shown in Listing 5.1.

Listing 5.1: Example of a structured JSON output demonstrating format adherence and analytical coherence.

```
1 {
```

```
2     "image_path": "data/FERET-Morphs/feret/morph_facemorpher
         /00955_960627_fb.png_00959_960627_fa.png.jpg",
3     "confidence_equal": 0.709219,
4     "reasoning": "```json\n{\n  \"final_decision\": {\n    \"
         overall_confidence_score\": 7250.872,\n    \"
         summary_of_findings\": \"The image exhibits several
         artifacts indicative of a potential morphing attack,
         falling into the High Suspicion range. The most prominent
         issues are inconsistencies in skin texture, particularly
         around the jawline and neck, and subtle blurring around
         the edges of the face. These, combined with minor
         asymmetries, suggest manipulation. While not definitive,
         the cumulative evidence strongly points towards a morph
         .\"\n  },\n  \"step_by_step_analysis\": {\n    \"
         step1_core_features\": {\n      \"score\": 6800.231,\n
              \"rationale\": \"The eyes appear relatively normal,
         but the lips exhibit a slight unnatural smoothness and a
         subtle lack of detail in the vermillion border. The nose
         appears slightly averaged, lacking the unique
         irregularities typically found in natural faces. This
         contributes to a high suspicion score.\"\n    },\n    \"
         step2_facial_geometry\": {\n      \"score\": 7500.555,\n
              \"rationale\": \"There's a noticeable asymmetry in
         the ears \u2013 the left ear appears slightly different in
          shape and position compared to the right. The jawline
         also seems somewhat averaged and lacks sharp definition,
         suggesting a blending of features. This asymmetry and ill-
         defined jawline raise significant concerns.\"\n    },\n
            \"step3_skin_texture\": {\n      \"score\": 7800.999,\n
               \"rationale\": \"The skin texture is overly smooth,
         particularly around the neck and jawline. There's a
         distinct lack of fine details like pores and micro-
         wrinkles in these areas, giving the skin a 'plastic'
         appearance. This is a strong indicator of manipulation.\"\
         n    },\n    \"step4_boundaries_and_edges\": {\n      \"
         score\": 6500.123,\n      \"rationale\": \"There's subtle
         blurring along the hairline and around the edges of the
         face, particularly where the face transitions into the
         background. This blurring suggests a potential seam or
         blending artifact, indicating manipulation.\"\n    },\n
            \"step5_lighting_and_color\": {\n      \"score\":
         6950.789,\n      \"rationale\": \"The lighting appears
```

```
        relatively consistent across the face, but there's a
        slight difference in the specular highlights in the eyes.
        The color tone is uniform, but the overall image lacks the
         subtle variations in color that are typically present in
        natural faces. This contributes to a moderate level of
        suspicion.\"\n    },\n    \"step6_identity_coherence\": {\
        n      \"score\": 7000.444,\n      \"rationale\": \"The
        face feels somewhat generic and lacks a strong, unique
        identity. The features seem averaged, and the overall
        impression is that the face doesn't quite 'hold together'
        as a convincing, real person. This lack of identity
        coherence adds to the suspicion.\"\n    }\n  }\n}\n```",
    "step_scores": {
      "step1_score": 6800.231,
      "step2_score": 7500.555,
      "step3_score": 7800.999,
      "step4_score": 6500.123,
      "step5_score": 6950.789,
      "step6_score": 7000.444
    }
  }
```

As demonstrated in the JSON output shown in Listing 5.1, each scored step is accompanied by a `rationale` field. This element serves a dual purpose: beyond providing a human-readable justification for the assigned score, it compels the LLM to articulate its reasoning. This process of verbalization appears to reinforce the model's ability to align its numerical output with the qualitative semantic score categories defined within the prompt.

An interesting component of the framework is the *identity coherence* step. This metric is not tangible in the sense that it cannot be directly measured by analyzing discrete pixels or geometric relationships. Instead, it prompts the model to perform a holistic evaluation, synthesizing the findings from the previous five steps to form a judgment on whether the constituent features coalesce into a convincing and singular identity. Interestingly, the inclusion of this abstract query was found to enhance the model's ability to detect concrete anomalies in the other, more tangible analytical steps, effectively helping it to round out and solidify its final decision. Excluding this question drops the model performance across the board.

5.1.2 Gemma-3 27B

Next, we analyze the zero-shot morphing attack detection performance of *Gemma-3 27B* across diverse datasets and morphing techniques. We examine its effectiveness against landmark-based, GAN-based, and diffusion-based attacks to establish baseline capabilities and identify technique-specific strengths and limitations.

Based on analysis of its zero-shot performance, the *Gemma-3 27B* model exhibits a foundational but unspecialized capability for morphing attack detection, achieving an overall average Equal Error Rate (EER) of **32.1%** (Table 5.2). Its effectiveness varies significantly depending on the specific morphing methodology employed.

Table 5.2: Zero-shot Equal Error Rate (%) for Gemma-3 27B across morphing techniques using structured forensic analysis. Results show combined performance and individual analytical step breakdowns (Step1: Core Features, Step2: Geometry, Step3: Skin Texture, Step4: Boundaries, Step5: Lighting, Step6: Identity Coherence). Lowest EER per row highlighted in bold.

Dataset	Morphing Technique	Combined	Step1	Step2	Step3	Step4	Step5	Step6
FERET	FaceMorpher	18.20	18.16	18.02	18.54	**17.51**	20.66	17.89
	OpenCV	19.62	19.48	19.34	19.86	**19.30**	21.60	20.52
	StyleGAN	40.94	39.31	39.17	**36.15**	41.71	39.51	41.09
FRGC	FaceMorpher	**32.19**	32.69	32.70	32.52	33.12	33.11	33.00
	OpenCV	43.55	43.53	43.75	43.49	**42.77**	43.86	44.46
	StyleGAN	57.06	**56.25**	56.46	56.47	56.78	57.39	57.43
FRLL	AMSL	25.10	**24.84**	25.79	27.52	26.04	27.34	30.55
	FaceMorpher	13.08	12.02	**11.94**	12.35	13.29	15.29	14.27
	OpenCV	13.33	13.33	**13.29**	13.95	13.54	15.13	15.50
	StyleGAN	27.39	27.76	**27.07**	27.72	29.51	28.58	28.21
	WebMorph	12.88	**12.06**	12.23	12.27	14.97	15.01	17.42
LMA	D	47.05	46.59	46.55	44.51	46.56	**44.01**	47.71
	PS	43.88	45.26	45.07	44.96	45.37	43.63	**43.03**
MIPGAN II	MIPGAN II	35.56	32.52	35.35	35.06	**31.63**	32.41	39.50
MorDiff	MorDiff	36.13	35.43	35.43	**34.05**	36.07	36.71	36.04
MorGAN	MorGAN	52.58	52.50	52.43	52.60	52.30	52.28	**52.01**
	LMA	52.87	52.89	52.77	**52.57**	53.22	52.67	53.17
Greedy	DIM	6.15	**5.92**	**5.92**	**5.92**	9.28	6.88	7.67
Average	–	**32.09**	**31.70**	**31.85**	**31.70**	**32.39**	**32.56**	**33.30**

Table 5.3: Zero-shot Attack Presentation Classification Error Rate (%) for Gemma-3 27B at fixed Bona Fide Presentation Classification Error Rate (BPCER) thresholds. Lower APCER values indicate better attack detection performance at each security threshold. Lowest APCER per row highlighted in bold.

Dataset	**Morphing Technique**	**@0.1%**	**@1%**	**@5%**	**@10%**	**@20%**	**@100%**
FERET	FaceMorpher	88.37	59.80	36.15	25.11	17.13	0.00
	OpenCV	88.14	60.66	34.83	26.83	20.27	0.00
	StyleGAN	100.00	96.79	78.55	69.00	58.99	0.00
FRGC	FaceMorpher	95.47	76.00	62.94	56.44	42.45	0.00
	OpenCV	97.10	86.41	76.21	72.19	58.31	0.00
	StyleGAN	100.00	100.00	98.41	97.28	87.18	0.00
FRLL	AMSL	99.80	96.08	81.11	59.50	34.91	0.00
	FaceMorpher	82.96	50.89	34.28	21.62	6.02	0.00
	OpenCV	72.23	37.01	27.63	19.32	8.73	0.00
	StyleGAN	99.98	96.96	87.52	71.06	35.83	0.00
	WebMorph	94.13	64.68	42.58	22.76	7.17	0.00
LMA	D	99.87	95.14	88.88	83.26	73.99	0.00
	PS	100.00	99.41	91.63	83.39	70.97	0.00
MIPGAN II	MIPGAN II	99.77	96.72	79.29	56.07	44.05	0.00
MorDiff	MorDiff	99.35	92.03	73.76	63.04	49.13	0.00
MorGAN	MorGAN	100.00	99.80	98.13	95.38	86.45	0.00
	LMA	100.00	99.71	99.10	95.76	85.67	0.00
Greedy	DIM	85.64	26.13	7.36	5.87	2.60	0.00
Average	–	**94.60**	**79.68**	**66.58**	**56.88**	**43.88**	**0.00**

The model performs best against traditional landmark-based morphing techniques, especially under controlled imaging conditions. On the *FRLL* dataset, it achieved relatively low EERs of **13.1%** against *FaceMorpher*, **13.3%** against *OpenCV*, and **12.9%** against *WebMorph*. However, performance on these same techniques degraded on more varied datasets like *FERET* (approx. 18–20% EER) and *FRGC* (32–44% EER), indicating a sensitivity to varied conditions and background noise.

Performance drops considerably against more advanced GAN-based attacks, which lack the obvious pixel-level artifacts of landmark-based methods. The model struggled to identify the subtle inconsistencies in these morphs, resulting in high error rates across all datasets. For instance, the EER on *StyleGAN*

Table 5.4: Zero-shot Bona Fide Presentation Classification Error Rate (%) for Gemma-3 27B at fixed Attack Presentation Classification Error Rate (APCER) thresholds. Lower BPCER values indicate better performance (fewer false alarms) at each security threshold. Lowest BPCER per row is highlighted in bold.

Dataset	**Morphing Technique**	**@0.1%**	**@1%**	**@5%**	**@10%**	**@20%**	**@100%**
FERET	FaceMorpher	97.42	73.12	46.24	31.09	14.76	**0.00**
	OpenCV	96.84	85.19	58.83	42.24	20.10	**0.00**
	StyleGAN	99.85	98.61	93.13	84.98	66.72	**0.00**
FRGC	FaceMorpher	98.69	93.22	76.46	64.86	45.84	**0.00**
	OpenCV	99.35	96.45	88.53	81.86	70.99	**0.00**
	StyleGAN	99.90	98.95	94.82	90.98	83.61	**0.00**
FRLL	AMSL	100.00	83.23	73.24	69.26	44.28	**0.00**
	FaceMorpher	80.88	53.33	22.06	14.22	11.39	**0.00**
	OpenCV	89.28	69.20	32.34	19.39	9.66	**0.00**
	StyleGAN	96.68	86.77	68.85	51.53	37.48	**0.00**
	WebMorph	96.30	71.61	27.25	14.22	12.25	**0.00**
LMA	D	99.67	96.71	88.35	81.70	69.12	**0.00**
	PS	99.61	97.42	83.01	75.58	61.97	**0.49**
MIPGAN II	MIPGAN II	99.77	88.77	76.62	62.93	52.35	**0.00**
MorDiff	MorDiff	98.87	98.36	82.45	66.42	50.48	**0.00**
MorGAN	MorGAN	99.26	97.77	94.75	89.85	84.07	**0.00**
	LMA	99.93	98.39	90.96	87.29	80.45	**0.00**
Greedy	DIM	94.61	38.24	12.99	2.45	1.47	**0.00**
Average	–	**97.05**	**84.74**	**67.27**	**57.27**	**45.39**	**0.03**

morphs was **27.4%** on *FRLL*, but climbed to **40.9%** on *FERET* and **57.1%** on *FRGC*. Similarly, it produced high EERs on other GAN-based datasets like *MIPGAN-II* (**35.6%**) and *MorGAN* (over **52%**).

The model's performance on modern diffusion-based morphs was inconsistent. While it struggled with the *MorDiff* technique, recording a high EER of **36.1%**, it achieved its best overall performance against *Greedy-DiM*, with an exceptionally low EER of **6.15%**.

From a practical standpoint, the model is not viable for deployment in its zero-shot state. The trade-off between security and user convenience is poor, as shown by its performance at fixed operating points (Table 5.3 and Table 5.4). To maintain a false rejection rate (BPCER) of 5%, the model would fail to detect

approximately **67%** of attacks (APCER). Conversely, to reliably detect 95% of morphing attacks (APCER of 5%), the system would incorrectly reject about **67%** of genuine users (BPCER). This demonstrates that while *Gemma-3* can perceive visual anomalies, it lacks the specialized calibration needed for reliable security applications without fine-tuning.

5.1.3 Qwen2.5-VL 32B

Following, we evaluate *Qwen2.5-VL*'s zero-shot morphing attack detection capabilities across multiple datasets and attack types. We assess its performance limitations and identify the factors contributing to its suboptimal detection accuracy.

Table 5.5: Zero-shot Equal Error Rate (EER, in %) for Qwen2.5-VL by dataset and morphing technique. The table includes breakdowns for each of the six analytical steps. Lower EER values indicate better overall detection performance. The lowest EER value in each row is highlighted in **bold**.

Dataset	Morphing Technique	Combined	Step1	Step2	Step3	Step4	Step5	Step6
FERET	FaceMorpher	**34.52**	35.07	36.32	37.19	37.70	35.35	35.23
	OpenCV	32.46	32.33	34.24	32.24	36.58	**32.00**	33.12
	StyleGAN	34.27	**33.84**	38.87	35.96	36.48	34.40	34.10
FRGC	FaceMorpher	42.41	40.27	41.25	42.40	43.29	**39.87**	41.64
	OpenCV	42.62	40.90	42.72	41.88	42.66	**40.37**	41.64
	StyleGAN	59.81	58.93	58.74	58.40	59.25	**57.96**	58.20
FRLL	AMSL	**44.13**	45.96	48.27	53.14	44.25	45.39	45.13
	FaceMorpher	43.01	43.74	44.73	42.24	**39.33**	42.27	40.03
	OpenCV	39.51	41.09	40.53	39.84	38.08	**37.22**	37.59
	StyleGAN	26.86	26.25	38.60	28.00	**26.17**	28.13	27.60
	WebMorph	41.06	42.15	44.01	50.79	38.86	**38.29**	39.18
LMA	D	45.40	**45.04**	48.18	46.24	46.51	46.33	45.78
	PS	47.43	**46.63**	47.08	46.76	46.68	46.92	48.64
MIPGAN II	MIPGAN II	20.75	20.44	21.72	**19.47**	**19.47**	20.28	20.14
MorDiff	MorDiff	45.91	46.55	**45.68**	46.48	46.56	**45.68**	46.68
MorGAN	MorGAN	48.63	51.14	49.17	**48.28**	50.11	49.85	50.63
	LMA	51.67	**49.25**	49.77	51.58	50.64	49.76	49.45
Greedy	DIM	**24.55**	25.35	30.11	24.85	25.09	26.38	25.82
Average	–	40.28	40.27	42.22	41.43	40.43	**39.80**	40.03

Table 5.6: Zero-shot Attack Presentation Classification Error Rate (APCER, in %) for Qwen2.5-VL at fixed Bona Fide Presentation Classification Error Rate (BPCER) thresholds. Lower APCER values indicate better attack detection performance at each security threshold.

Dataset	**Morphing Technique**	**@0.1%**	**@1%**	**@5%**	**@10%**	**@20%**	**@100%**
FERET	FaceMorpher	99.98	95.09	63.33	55.87	46.57	0.00
	OpenCV	99.79	96.41	55.01	46.03	39.75	0.00
	StyleGAN	99.38	81.66	53.50	51.65	43.85	0.00
FRGC	FaceMorpher	99.64	96.06	92.57	85.33	73.60	0.10
	OpenCV	99.02	95.67	91.77	82.83	67.63	0.10
	StyleGAN	98.32	96.59	94.93	90.77	82.88	0.10
FRLL	AMSL	99.97	99.75	99.36	90.95	70.00	13.70
	FaceMorpher	99.96	99.55	84.61	77.38	60.34	0.41
	OpenCV	99.86	99.36	77.72	69.53	55.94	0.57
	StyleGAN	98.81	88.09	54.26	47.78	30.46	0.00
	WebMorph	99.97	99.72	91.73	83.97	62.75	3.03
LMA	D	99.47	97.19	87.81	83.04	70.50	0.74
	PS	99.83	99.41	94.57	92.38	86.41	0.00
MIPGAN II	MIPGAN II	44.20	42.53	35.85	32.02	20.94	0.00
MorDiff	MorDiff	98.58	92.89	83.22	77.46	71.20	0.08
MorGAN	MorGAN	100.00	100.00	96.42	90.17	82.95	0.35
	LMA	100.00	99.30	96.11	90.88	82.05	0.12
Greedy	DIM	99.31	93.06	32.56	29.97	27.60	0.00
Average	–	**96.45**	**92.91**	**76.96**	**71.00**	**59.75**	**1.07**

While the Qwen2.5-VL model demonstrates a basic ability to distinguish between genuine and morphed images, its performance falls significantly below the state-of-the-art in morphing attack detection.The model consistently yields EERs between 20% and 60%. This indicates that its detection capability is often only marginally better than random guessing. The model's effectiveness varies depending on the morphing technique used. A detailed breakdown is shown in Table 5.5.

A deeper look at the EER results reveals a clear pattern: the model struggles with nearly all categories of morphing attacks. Against **classical landmark-based morphs** such as OpenCV and FaceMorpher, Qwen2.5-VL performs very poorly. For example, it achieves 39.51% EER on FRLL-OpenCV, 32.46% on FERET-OpenCV, 42.41% on FRGC-FaceMorpher. These results indicate that

Table 5.7: Zero-shot Bona Fide Presentation Classification Error Rate (BPCER, in %) for Qwen2.5-VL at fixed Attack Presentation Classification Error Rate (APCER) thresholds. Lower BPCER values indicate better performance (fewer false alarms) at each security threshold.

Dataset	**Morphing Technique**	**@0.1%**	**@1%**	**@5%**	**@10%**	**@20%**	**@100%**
FERET	FaceMorpher	97.81	93.00	82.97	74.35	62.77	0.00
	OpenCV	95.39	93.29	81.43	74.06	54.71	0.00
	StyleGAN	99.82	97.99	93.68	88.49	63.87	0.00
FRGC	FaceMorpher	99.81	95.58	82.10	75.07	57.32	0.09
	OpenCV	99.40	95.72	86.75	77.32	59.78	0.00
	StyleGAN	99.81	98.55	94.47	87.88	82.76	0.00
FRLL	AMSL	100.00	100.00	100.00	100.00	94.08	0.00
	FaceMorpher	100.00	98.59	95.22	89.15	75.40	0.00
	OpenCV	100.00	98.67	94.61	87.06	72.24	0.00
	StyleGAN	100.00	100.00	97.26	93.14	51.72	0.00
	WebMorph	100.00	100.00	97.92	91.66	78.24	0.00
LMA	D	97.49	97.49	89.94	85.97	75.70	0.00
	PS	99.79	99.33	90.86	87.17	75.53	0.33
MIPGAN II	MIPGAN II	99.21	77.70	30.56	25.71	20.71	0.00
MorDiff	MorDiff	99.19	97.24	89.96	84.15	76.86	0.00
MorGAN	MorGAN	100.00	98.07	93.48	88.46	79.27	0.00
	LMA	99.92	99.09	95.25	91.47	81.84	0.00
Greedy	DIM	99.51	97.25	87.12	74.02	41.54	0.00
Average	–	**99.29**	**96.53**	**87.98**	**81.95**	**66.91**	**0.02**

the model is incapable of reliably identifying even the relatively well-studied artifacts of older morphing approaches, such as ghosting, blurred contours, and unnatural skin textures.

The model is equally ineffective against **GAN-based morphs** such as StyleGAN, MIPGAN, and MorGAN. Performance is inconsistent: while it achieves its best result of 20.75% EER on the MIPGAN dataset, its detection rates collapse on other benchmarks, reaching 26.86% on FRLL-StyleGAN, 34.27% on FERET-StyleGAN, 48.63% on MorGAN, and 59.81% on FRGC-StyleGAN. This demonstrates that Qwen2.5-VL is often completely fooled by high-fidelity GAN morphs, which lack the obvious visual cues of older morphing methods. The model fails to detect systemic inconsistencies of GAN synthesis such as artificial symmetry or missing fine-grained details like skin pores.

Finally, the model is also ineffective against **modern diffusion-based morphs**. On Greedy-DiM it records a 24.55% EER, while on MorDiff it fails with 45.91% EER. These results confirm that Qwen2.5-VL cannot cope with next-generation morphing attacks that are specifically designed to be artifact-free and visually seamless.

The trade-off analysis reported in Table 5.6 and Table 5.7 further illustrates the model's inability to approach state-of-the-art results. In the APCER@BPCER analysis, when the system is tuned to maintain a low false rejection rate (for example, fixing the BPCER at 1% or 5%), the Attack Presentation Classification Error Rate (APCER) approaches 100% for most datasets. This means that nearly all morphing attacks bypass detection, showing that the model cannot simultaneously maintain usability and robustness. Conversely, the BPCER@APCER analysis shows that when the system is tuned to reliably detect morphing attacks (for example, fixing the APCER at 1% or 5%), the BPCER rises above 90% in many cases. This implies that an unacceptably high proportion of genuine users would be falsely classified as attackers.

In summary, the Qwen2.5-VL model falls short of state-of-the-art performance for morphing attack detection. Its high EER values across classical, GAN-based, and diffusion-based morphs, together with its extreme failure to balance APCER and BPCER at practical operating points, confirm that it is not a viable candidate for advancing morph detection performance.

5.1.4 Llama-4-Scout 17B

In this section, we examine *Llama-4-Scout*'s zero-shot morphing attack detection performance and analyze the underlying causes of its poor detection capabilities, including issues with information synthesis and holistic reasoning.

In its zero-shot evaluation, the *Llama-4-Scout* model demonstrates a poor and inconsistent capability for morphing attack detection, achieving an overall average Equal Error Rate (EER) of **44.6%** (Table 5.8). The model's performance reveals a flaw in its reasoning process: its holistic, combined judgment is often less accurate than its analysis of specific, isolated visual components. This suggests

Table 5.8: Zero-shot Equal Error Rate (EER, in %) for Llama-4-Scout by dataset and morphing technique. The table includes breakdowns for each of the six analytical steps. Lower EER values indicate better overall detection performance. The lowest EER value in each row is highlighted in **bold**.

Dataset	Morphing Technique	Combined	Step1	Step2	Step3	Step4	Step5	Step6
FRLL	AMSL	49.63	44.05	37.09	39.78	**36.59**	43.01	42.71
	FaceMorpher	41.50	**21.02**	36.67	29.74	23.47	39.87	44.76
	OpenCV	37.63	21.89	32.71	22.59	**17.71**	39.88	45.49
	StyleGAN	**47.59**	55.91	55.12	70.02	56.00	48.37	50.29
	WebMorph	39.22	26.84	29.81	29.02	**25.20**	44.39	44.51
Greedy	DIM	**49.93**	57.05	50.86	50.99	51.00	52.94	50.47
MIPGAN II	MIPGAN II	46.58	60.38	61.59	55.45	55.20	57.50	**43.66**
Average	–	44.58	41.02	43.41	42.51	**37.88**	46.57	45.98

a failure in synthesizing information, where the model can identify certain low-level artifacts but becomes confused when attempting to form a comprehensive conclusion. This could be attributed to the quantization of the model, where the model's ability to analyze individual components remains intact while its holistic, combined judgment becomes less accurate than the analysis of isolated visual or reasoning components [87].

A detailed analysis of the multi-step evaluation framework shows that the final "Combined" EER of **44.6%** is notably worse than the performance of several individual steps. The most effective single analytical instruction was Step 4 (*Boundary and Edge Analysis*), which achieved a significantly better average EER of **37.9%**. This step prompts the model to check for common blending artifacts at the face perimeter, such as the jawline and hairline. Its relative success indicates that the model is most effective when given a concrete, localized task that directly corresponds to the known weaknesses of traditional morphing techniques. The model's limited proficiency is almost entirely concentrated on these landmark-based morphs, where Step 4 recorded its best result of **17.7%** EER on the *FRLL-OpenCV* set.

However, the model's performance collapses against modern, representation-space attacks that do not produce such obvious edge artifacts. On GAN-based (*StyleGAN*, *MIPGAN-II*) and diffusion-based (*Greedy-DiM*) morphs, EERs consistently exceeded **45%**, indicating performance close to random chance. The fail-

Table 5.9: Zero-shot Attack Presentation Classification Error Rate (APCER, in %) for Llama-4-Scout at fixed Bona Fide Presentation Classification Error Rate (BPCER) thresholds. Lower APCER values indicate better attack detection performance at each security threshold.

Dataset	**Morphing Technique**	**@0.1%**	**@1%**	**@5%**	**@10%**	**@20%**	**@100%**
FRLL	AMSL	99.99	99.03	98.27	96.96	96.64	0.00
	FaceMorpher	98.78	91.77	86.43	84.01	83.72	0.08
	OpenCV	99.78	85.55	66.37	62.15	61.67	0.00
	StyleGAN	100.00	100.00	99.87	98.23	76.81	0.08
	WebMorph	99.82	95.46	83.88	81.68	81.24	0.00
Greedy	DIM	99.98	99.78	98.87	97.16	82.01	0.00
MIPGAN II	MIPGAN II	100.00	99.86	99.77	99.01	68.29	0.00
Average	–	**99.76**	**95.92**	**90.49**	**88.46**	**78.63**	**0.02**

ure of the final "Combined" score suggests that more abstract or holistic prompts, such as Step 5 (*Lighting Consistency*) and Step 6 (*Identity Coherence*), introduce noise and detract from the more reliable signals found in Step 4. This indicates that *Llama-4-Scout* struggles to weigh and integrate different pieces of visual evidence, leading to a flawed final judgment.

From a practical deployment standpoint, the model is entirely unsuitable. The trade-off analysis at fixed operating points (Table 5.9 and Table 5.10) confirms this deficiency. To limit the rejection of genuine users to 5% (BPCER), the system would fail to detect over **90%** of morphing attacks (APCER). Conversely, to reliably detect 95% of attacks, it would incorrectly reject nearly **90%** of legitimate users. This demonstrates that *Llama-4-Scout*, in its zero-shot state, cannot perform the complex forensic reasoning required for this task.

5.1.5 Mistral Small 3.1 24B

In this section, we evaluate *Mistral Small 3.1*'s zero-shot morphing attack detection performance across multiple datasets and morphing techniques. We analyze its systematic bias toward classifying morphed images as authentic and examine why it achieves near-random detection capability.

Table 5.10: Zero-shot Bona Fide Presentation Classification Error Rate (BPCER, in %) for Llama-4-Scout at fixed Attack Presentation Classification Error Rate (APCER) thresholds. Lower BPCER values indicate better performance (fewer false alarms) at each security threshold.

Dataset	**Morphing Technique**	**@0.1%**	**@1%**	**@5%**	**@10%**	**@20%**	**@100%**
FRLL	AMSL	100.00	100.00	88.24	86.42	72.45	0.00
	FaceMorpher	100.00	100.00	88.24	77.92	55.88	0.00
	OpenCV	100.00	100.00	88.24	75.18	53.01	0.00
	StyleGAN	100.00	100.00	95.10	92.04	87.51	0.00
	WebMorph	94.34	88.24	80.39	72.88	53.43	0.00
Greedy	DIM	99.89	98.92	94.61	88.36	82.83	0.00
MIPGAN II	MIPGAN II	99.53	98.11	94.79	93.55	88.56	0.00
Average	–	**99.11**	**97.90**	**89.94**	**83.76**	**70.52**	**0.00**

In a zero-shot evaluation, the *Mistral Small 3.1 24B* model proves to be entirely ineffective for morphing attack detection, with its performance consistently approaching that of random chance. The model's overall average Equal Error Rate (EER) of **48.4%** (Table 5.11) underscores a near-complete inability to distinguish between genuine and morphed images. This comprehensive failure stems from the model's strong tendency to confidently misclassify manipulated images as authentic. In its responses, the model frequently assigned suspicion scores near 1,000, a range that, according to the provided analytical framework, signifies a high likelihood of the image being a "bona fide" picture with no morphing artifacts.

This inherent bias towards a "genuine" classification holds true across all morphing techniques. The model failed to detect even the more overt artifacts of traditional landmark-based methods, yielding poor EERs ranging from **37.3%** to as high as **53.3%**. Its performance degraded even further against advanced GAN-based attacks, which lack obvious pixel-level seams; it recorded a **52.1%** EER against *StyleGAN* and its worst result of **55.6%** against *MorGAN*. Similarly, *Mistral* was completely confounded by modern, high-realism diffusion morphs, registering a **50.5%** EER on the challenging *Greedy-DiM* dataset.

From a practical standpoint, the model is entirely non-viable for security applications. Analysis at fixed operating points (Table 5.12 and Table 5.13) reveals

Table 5.11: Zero-shot Equal Error Rate (EER, in %) for Mistral Small 3.1 by dataset and morphing technique. The table includes breakdowns for each of the six analytical steps. Lower EER values indicate better overall detection performance. The lowest EER value in each row is highlighted in **bold**.

Dataset	Morphing Technique	Combined	Step1	Step2	Step3	Step4	Step5	Step6
FERET	FaceMorpher	53.07	57.15	49.84	50.87	50.02	**44.94**	58.33
	OpenCV	53.26	55.26	47.73	51.32	47.19	**45.79**	58.14
	StyleGAN	48.57	48.07	**47.25**	55.07	49.36	51.89	49.16
FRGC	FaceMorpher	49.20	53.45	**48.12**	49.74	48.34	51.43	52.90
	OpenCV	52.82	50.76	51.96	47.72	**46.79**	52.85	57.77
	StyleGAN	50.45	53.87	**41.93**	54.20	52.44	54.66	49.93
FRLL	AMSL	**42.13**	51.24	43.60	49.88	48.72	46.83	50.49
	FaceMorpher	**37.29**	48.99	38.80	53.56	45.57	43.13	46.61
	OpenCV	40.24	51.97	**39.55**	55.46	44.35	46.10	47.44
	StyleGAN	52.06	55.40	**41.55**	53.37	57.29	51.29	52.11
	WebMorph	**40.78**	50.33	41.02	49.97	47.18	44.30	49.41
LMA	D	49.01	50.29	47.09	**44.10**	48.34	49.82	49.99
	PS	51.50	52.66	51.68	50.50	48.91	**46.39**	52.08
MIPGAN II	MIPGAN II	50.24	52.93	**40.68**	47.03	55.40	52.03	56.84
MorDiff	MorDiff	47.77	49.22	46.98	48.60	47.01	**45.70**	46.76
MorGAN	MorGAN	55.60	53.51	54.71	55.60	**52.91**	54.46	55.65
	LMA	46.88	46.99	**46.78**	47.18	48.21	47.03	47.14
Greedy	DIM	50.49	50.47	**47.81**	51.17	52.47	51.84	50.23
Average	–	48.41	51.81	**45.95**	50.85	49.47	48.92	51.72

an unusable trade-off; to limit false rejections of genuine users to a reasonable 5%, the system would consequently fail to detect **95%** of morphing attacks. This demonstrates that in its unadapted state, *Mistral* lacks the specialized forensic judgment for this task, defaulting to an incorrect assessment of authenticity regardless of the visual evidence.

5.1.6 Comparative Zero-Shot Performance Analysis

In this section, we compare zero-shot performance across *Gemma-3*, *Qwen2.5-VL*, *Mistral Small 3.1*, and *Llama-4-Scout*. Table 5.14 reports combined Equal Error Rates (EER) averaged within each dataset group. Tables 5.15 and 5.16 summarize operating-point behavior at standardized thresholds, i.e., **APCER@BPCER = 1%, 5%** and **BPCER@APCER = 1%, 5%**, respectively. Missing entries ("–")

Table 5.12: Zero-shot Attack Presentation Classification Error Rate (APCER, in %) for Mistral Small 3.1 at fixed Bona Fide Presentation Classification Error Rate (BPCER) thresholds. Lower APCER values indicate better attack detection performance at each security threshold.

Dataset	**Morphing Technique**	**@0.1%**	**@1%**	**@5%**	**@10%**	**@20%**	**@100%**
FERET	FaceMorpher	99.87	99.27	96.03	92.11	81.48	0.00
	OpenCV	99.73	98.46	91.30	85.84	76.66	0.00
	StyleGAN	99.60	98.96	97.16	93.97	81.04	0.00
FRGC	FaceMorpher	99.97	99.45	97.57	92.16	79.40	0.00
	OpenCV	100.00	99.48	97.50	94.38	86.20	0.00
	StyleGAN	100.00	99.92	99.27	96.12	83.19	0.00
FRLL	AMSL	99.93	99.31	95.08	83.96	67.31	0.05
	FaceMorpher	99.81	98.10	90.56	77.67	59.31	0.00
	OpenCV	99.83	98.30	94.19	81.05	62.36	0.00
	StyleGAN	99.99	99.93	99.13	90.82	80.77	0.49
	WebMorph	99.83	98.30	90.53	78.37	60.29	0.00
LMA	D	100.00	99.77	92.43	82.17	73.48	0.00
	PS	100.00	99.23	96.18	91.54	77.89	0.00
MIPGAN II	MIPGAN II	99.93	99.24	90.92	84.72	74.66	0.00
MorDiff	MorDiff	99.51	98.42	93.05	84.51	72.00	0.00
MorGAN	MorGAN	100.00	99.75	96.71	92.64	86.06	0.00
	LMA	100.00	99.50	96.58	90.99	79.05	0.00
Greedy	DIM	99.80	98.99	96.49	93.36	84.80	0.00
Average	–	**99.88**	**99.13**	**95.04**	**88.13**	**75.89**	**0.03**

indicate that no results were available for that (model, dataset) pair.

A comparative analysis of the zero-shot performance of the four multimodal large language models (MLLMs) reveals a distinct hierarchy of capability, with *Gemma-3 27B* emerging as the clear frontrunner, while other models exhibit significant limitations. As shown in Table 5.14, *Gemma-3* achieved the best overall performance with an average Equal Error Rate (EER) of 32.09%, followed by *Qwen2.5-VL* at 40.28%. *Llama-4-Scout* and *Mistral Small 3.1* were largely ineffective, with average EERs of 44.58% and 48.41%, respectively, indicating performance often at or near random chance.

Gemma-3's superior performance is primarily driven by its unique ability to handle both traditional and modern morphing attacks. It demonstrated strong

Table 5.13: Zero-shot Bona Fide Presentation Classification Error Rate (BPCER, in %) for Mistral Small 3.1 at fixed Attack Presentation Classification Error Rate (APCER) thresholds. Lower BPCER values indicate better performance (fewer false alarms) at each security threshold.

Dataset	**Morphing Technique**	**@0.1%**	**@1%**	**@5%**	**@10%**	**@20%**	**@100%**
FERET	FaceMorpher	99.96	99.43	97.03	94.27	89.73	0.00
	OpenCV	99.78	99.43	97.13	94.36	88.06	0.00
	StyleGAN	100.00	99.25	91.48	88.05	80.11	0.00
FRGC	FaceMorpher	98.90	97.83	94.32	89.09	80.57	0.00
	OpenCV	99.63	98.35	95.89	92.09	84.43	0.00
	StyleGAN	99.85	99.42	91.15	87.72	84.00	0.00
FRLL	AMSL	99.81	98.53	89.95	81.37	62.80	0.00
	FaceMorpher	99.89	97.11	88.06	79.46	58.69	0.00
	OpenCV	99.89	97.51	88.12	81.37	64.85	0.00
	StyleGAN	100.00	100.00	94.61	91.10	85.22	0.00
	WebMorph	99.51	98.26	94.58	82.35	64.24	0.00
LMA	D	99.85	99.24	96.98	92.03	82.01	0.00
	PS	99.97	99.14	95.62	91.97	83.38	0.49
MIPGAN II	MIPGAN II	99.29	99.29	91.45	85.39	74.63	0.00
MorDiff	MorDiff	99.51	99.25	93.21	85.28	73.19	0.00
MorGAN	MorGAN	99.80	99.00	95.92	92.53	86.79	0.00
	LMA	99.26	98.17	91.41	83.33	73.09	0.00
Greedy	DIM	100.00	98.53	94.61	90.39	80.73	0.00
Average	–	**99.72**	**98.76**	**93.42**	**87.90**	**77.58**	**0.03**

proficiency in detecting landmark-based morphs, achieving EERs around 13% on the *FRLL* dataset's *FaceMorpher*, *OpenCV*, and *WebMorph* subsets. Most notably, it was the only model to effectively identify the sophisticated diffusion-based *Greedy-DiM* attacks, achieving an exceptionally low EER of 6.15% (Table 5.14). This suggests that *Gemma-3*'s visual reasoning is capable of identifying the subtle, consistent artifacts present in both simple and highly realistic manipulations, although it still struggled with certain GAN-based morphs such as *FRGC-StyleGAN* (57.1% EER).

Qwen2.5-VL positioned itself as a distant second. While generally outperformed by *Gemma-3*, it showed surprising strength against specific GAN-based attacks, achieving the best EER on the challenging *MIPGAN II* dataset (20.75%) (Table 5.14) and on several *StyleGAN* subsets. This suggests that its architec-

Table 5.14: Zero-shot Equal Error Rate (EER, in %) by dataset and morphing technique, compared across all models. Lower values indicate better performance. The best-performing model for each technique is highlighted in **bold**.

Dataset	Technique	Gemma-3 27B	Qwen2.5-VL 32B	Llama-4-Scout 17B	Mistral Small 3.1 24B
FERET	FaceMorpher	**18.20**	34.52	—	53.07
	OpenCV	**19.62**	32.46	—	53.26
	StyleGAN	40.94	**34.27**	—	48.57
FRGC	FaceMorpher	**32.19**	42.41	—	49.20
	OpenCV	43.55	**42.62**	—	52.82
	StyleGAN	57.06	59.81	—	**50.45**
FRLL	AMSL	**25.10**	44.13	49.63	42.13
	FaceMorpher	**13.08**	43.01	41.50	37.29
	OpenCV	**13.33**	39.51	37.63	40.24
	StyleGAN	27.39	**26.86**	47.59	52.06
	WebMorph	**12.88**	41.06	39.22	40.78
LMA	D	47.05	**45.40**	—	49.01
	PS	**43.88**	47.43	—	51.50
MIPGAN II	MIPGAN II	35.56	**20.75**	46.58	50.24
MorDiff	MorDiff	**36.13**	45.91	—	47.77
MorGAN	MorGAN	52.58	**48.63**	—	55.60
	LMA	52.87	51.67	—	**46.88**
Greedy	DIM	**6.15**	24.55	49.93	50.49
Average	–	**32.09**	40.28	44.58	48.41

tural design may possess a different inductive bias that is more attuned to certain types of synthetic image generation, even if its overall forensic capability is lower.

In contrast, *Llama-4-Scout* and *Mistral Small 3.1* proved unsuitable for the task. *Llama-4-Scout*'s poor performance appears linked to a failure in synthesizing information from its analytical steps, while *Mistral* consistently defaulted to a "bona fide" classification, rendering it unable to detect nearly all forms of manipulation.

Despite *Gemma-3*'s relative success, the analysis of performance at fixed operating points (Tables 5.15 and 5.16) confirms that none of the models are viable for practical deployment in a zero-shot configuration. Even *Gemma-3* exhibits an unacceptable trade-off between security and usability; at a 5% BPCER (a threshold for low user friction), it would still allow approximately 67% of attacks to pass undetected (APCER). The other models perform far worse, with their APCER values at the same BPCER threshold exceeding 71% for *Qwen2.5-VL* and 90% for *Llama* and *Mistral*.

Table 5.15: Zero-shot Attack Presentation Classification Error Rate (APCER, in %) at fixed 1% and 5% Bona Fide Presentation Classification Error Rate (BPCER) thresholds, compared across all models. Lower values indicate better attack detection performance. The best-performing model for each operating point and technique is highlighted in **bold**.

Dataset	**Technique**	**Gemma-3 27B**		**Qwen2.5-VL 32B**		**Llama-4-Scout 17B**		**Mistral Small 3.1 24B**	
		1%	**5%**	**1%**	**5%**	**1%**	**5%**	**1%**	**5%**
FERET	FaceMorpher	**59.80**	**36.15**	95.09	63.33	—	—	99.27	96.03
	OpenCV	**60.66**	**34.83**	96.41	55.01	—	—	98.46	91.30
	StyleGAN	96.79	78.55	**81.66**	**53.50**	—	—	98.96	97.16
FRGC	FaceMorpher	**76.00**	**62.94**	96.06	92.57	—	—	99.45	97.57
	OpenCV	**86.41**	**76.21**	95.67	91.77	—	—	99.48	97.50
	StyleGAN	100.00	98.41	**96.59**	**94.93**	—	—	99.92	99.27
FRLL	AMSL	**96.08**	**81.11**	99.75	99.36	99.03	98.27	99.31	95.08
	FaceMorpher	**50.89**	**34.28**	99.55	84.61	91.77	86.43	98.10	90.56
	OpenCV	**37.01**	**27.63**	99.36	77.72	85.55	66.37	98.30	94.19
	StyleGAN	96.96	87.52	**88.09**	**54.26**	100.00	99.87	99.93	99.13
	WebMorph	**64.68**	**42.58**	99.72	91.73	95.46	83.88	98.30	90.53
LMA	D	**95.14**	88.88	97.19	**87.81**	—	—	99.77	92.43
	PS	99.41	**91.63**	99.41	94.57	—	—	**99.23**	96.18
MIPGAN II	MIPGAN II	96.72	79.29	**42.53**	**35.85**	99.86	99.77	99.24	90.92
MorDiff	MorDiff	**92.03**	**73.76**	92.89	83.22	—	—	98.42	93.05
MorGAN	MorGAN	99.80	98.13	100.00	**96.42**	—	—	**99.75**	96.71
	LMA	99.71	99.10	**99.30**	**96.11**	—	—	99.50	96.58
Greedy	DIM	**26.13**	**7.36**	93.06	32.56	99.78	98.87	98.99	96.49
Average	–	**79.68**	**66.58**	92.91	76.96	95.92	90.49	99.13	95.04

The zero-shot evaluation results indicate that while performance is not competitive with state-of-the-art detectors [1], multimodal large language models (MLLMs) possess an inherent capability for morphing attack detection without task-specific training. Given the substantial computational and time requirements for *LoRA* fine-tuning of multimodal large language models, we focused our adaptation efforts on a single architecture. The clearly superior performance of *Gemma-3* compared to other evaluated models (32.09% vs. 40.28% average EER) made it the obvious choice for specialized training. These findings suggest that, while the choice of model architecture is a significant factor in establishing baseline performance, substantial domain-specific adaptation is required to achieve the detection accuracy necessary to rival the current state-of-the-art MAD models.

Table 5.16: Zero-shot Bona Fide Presentation Classification Error Rate (BPCER, in %) at fixed 1% and 5% Attack Presentation Classification Error Rate (APCER) thresholds, compared across all models. Lower values indicate better performance (fewer false alarms). The best-performing model for each operating point and technique is highlighted in **bold**.

Dataset	**Technique**	**Gemma-3 27B**		**Qwen2.5-VL**		**Llama-4-Scout 17B**		**Mistral Small 3.1 24B**	
		1%	**5%**	**1%**	**5%**	**1%**	**5%**	**1%**	**5%**
FERET	FaceMorpher	**73.12**	**46.24**	93.00	82.97	—	—	99.43	97.03
	OpenCV	**85.19**	**58.83**	93.29	81.43	—	—	99.43	97.13
	StyleGAN	98.61	93.13	**97.99**	93.68	—	—	99.25	**91.48**
FRGC	FaceMorpher	**93.22**	**76.46**	95.58	82.10	—	—	97.83	94.32
	OpenCV	96.45	88.53	**95.72**	**86.75**	—	—	98.35	95.89
	StyleGAN	98.95	94.82	**98.55**	94.47	—	—	99.42	**91.15**
FRLL	AMSL	**83.23**	**73.24**	100.00	100.00	100.00	88.24	98.53	89.95
	FaceMorpher	**53.33**	**22.06**	98.59	95.22	100.00	88.24	97.11	88.06
	OpenCV	**69.20**	**32.34**	98.67	94.61	100.00	88.24	97.51	88.12
	StyleGAN	**86.77**	**68.85**	100.00	97.26	100.00	95.10	100.00	94.61
	WebMorph	**71.61**	**27.25**	100.00	97.92	88.24	80.39	98.26	94.58
LMA	D	**96.71**	**88.35**	97.49	89.94	—	—	99.24	96.98
	PS	**97.42**	**83.01**	99.33	90.86	—	—	99.14	95.62
MIPGAN II	MIPGAN II	88.77	76.62	**77.70**	**30.56**	98.11	94.79	99.29	91.45
MorDiff	MorDiff	98.36	**82.45**	**97.24**	89.96	—	—	99.25	93.21
MorGAN	MorGAN	**97.77**	94.75	98.07	**93.48**	—	—	99.00	95.92
	LMA	98.39	**90.96**	99.09	95.25	—	—	**98.17**	91.41
Greedy	DIM	**38.24**	**12.99**	97.25	87.12	98.92	94.61	98.53	94.61
Average		**84.74**	**67.27**	96.53	87.98	97.90	89.94	98.76	93.42

5.2 LoRA Fine-Tuned Gemma-3 12B resutls

In this section, we evaluate the performance improvements achieved through LoRA fine-tuning of *Gemma-3 12B* for morphing attack detection. We compare the fine-tuned model against our zero-shot experiments and a classification head-only baseline to demonstrate the effectiveness of parameter-efficient adaptation.

The evaluation results demonstrate the effectiveness of LoRA fine-tuning in enhancing morphing attack detection while validating that the base model retains inherent discriminative ability without adaptation. Table 5.17 presents a comprehensive comparison between three configurations: zero-shot evaluation of the *Gemma-3 27B* model, classification head-only baseline using the frozen *Gemma-3 12B* model, and LoRA fine-tuning of the *Gemma-3 12B* model.

Table 5.17: Equal Error Rate (EER) comparison of *Gemma-3* models: Zero-Shot (Combined), Classification Head only fine-tuning, and LoRA fine-tuning. All values are reported in percentages with two decimal precision. The final column (Δ) shows the difference in EER between LoRA and ClassHead Only (LoRA − HeadOnly). Negative Δ indicates that LoRA achieved a lower error rate (improvement), while positive Δ indicates worse performance. Best per row is highlighted in **bold**.

Dataset	**Technique**	**Zero-Shot**	**ClassHead Only**	**LoRA Fine-Tuned**	**Δ (LoRA − HeadOnly)**
FERET	FaceMorpher	18.20	13.80	**9.30**	-4.50
	OpenCV	19.62	11.73	**7.40**	-4.33
	StyleGAN	40.94	**27.40**	34.97	+7.57
FRGC	FaceMorpher	32.19	24.48	**4.98**	-19.50
	OpenCV	43.55	29.72	**9.23**	-20.49
	StyleGAN	57.06	40.82	**17.32**	-23.50
FRLL	AMSL	25.10	48.62	**12.74**	-35.88
	FaceMorpher	13.08	18.15	**0.49**	-17.66
	OpenCV	13.33	6.38	**1.47**	-4.91
	StyleGAN	27.39	16.72	**6.83**	-9.89
	WebMorph	12.88	21.47	**2.37**	-19.10
LMA	D	47.05	31.56	**18.90**	-12.66
	PS	43.88	39.74	**28.28**	-11.46
MIPGAN II	MIPGAN II	35.56	31.67	**17.92**	-13.75
MorDiff	MorDiff	36.13	24.88	**17.33**	-7.55
MorGAN	MorGAN	52.58	46.69	**45.26**	-1.43
	LMA	52.87	50.00	**23.06**	-26.94
Greedy	DIM	**6.15**	11.24	11.78	+0.54
Average		32.09	27.50	**14.98**	-12.52

The classification head-only baseline serves a critical methodological purpose by isolating the pre-trained model's inherent capability to distinguish morphs from bona fide images without any parameter updates. This configuration achieved an average Equal Error Rate (EER) of 27.50%, demonstrating that the frozen *Gemma-3 12B* model possesses substantial discriminative knowledge despite being 15 billion parameters smaller than the zero-shot variant. Unlike zero-shot evaluation, where text generation can influence absolute performance metrics, the classification head directly leverages representations from the final language model layer, providing a cleaner assessment of the model's latent forensic capabilities.

LoRA fine-tuning reduced the average EER to 14.98%, representing a 45.5% relative improvement over the frozen baseline. This substantial gain, achieved

Table 5.18: **Attack Presentation Classification Error Rate (APCER)** at fixed **Bona Fide Presentation Classification Error Rate (BPCER)** thresholds for *Gemma-3* model variants. Comparison includes *Zero-Shot (27B)*, *Classification Head-Only baseline (12B frozen)*, and *LoRA fine-tuned (12B)* configurations across five operating points (0.1%, 1%, 5%, 10%, 20% BPCER). Lower **APCER values indicate better attack detection** at each security threshold. Best results per operating point are highlighted in **bold**.

Dataset	**Technique**	**@0.1%**			**@1%**			**@5%**			**@10%**			**@20%**		
		ZS	HO	LoRA	ZS	HO	LoRA	ZS	HO	LoRA	ZS	HO	LoRA	ZS	HO	LoRA
FERET	FaceMorpher	88.37	61.49	**50.37**	59.80	31.02	**27.17**	36.15	18.15	**14.18**	25.11	15.69	**7.94**	17.13	11.53	**4.91**
	OpenCV	88.14	65.30	**53.17**	60.66	33.27	**24.85**	34.83	17.58	**10.02**	26.83	12.48	**6.43**	20.27	7.37	**3.21**
	StyleGAN	100.00	99.81	**98.79**	96.79	95.27	**92.29**	78.55	74.18	**73.16**	69.00	**56.33**	58.03	58.99	**35.92**	48.02
FRGC	FaceMorpher	95.47	65.56	**48.10**	76.00	51.49	**19.82**	62.94	39.80	**4.98**	56.44	34.75	**2.70**	42.45	28.92	**1.14**
	OpenCV	97.10	81.02	**72.93**	86.41	66.01	**37.43**	76.21	52.56	**14.11**	72.19	46.37	**8.40**	58.31	39.32	**3.53**
	StyleGAN	100.00	100.00	**99.02**	100.00	99.59	**82.89**	98.41	91.98	**47.62**	97.28	81.62	**33.34**	87.18	61.62	**15.25**
FRLL	AMSL	99.80	99.96	**87.73**	96.08	99.46	**62.78**	81.11	95.99	**24.94**	59.50	87.91	**15.58**	34.91	77.94	**5.24**
	FaceMorpher	82.96	56.95	**4.89**	50.89	35.15	**0.48**	34.28	25.99	**0.16**	21.62	21.34	**0.16**	6.02	17.61	**0.08**
	OpenCV	72.23	28.75	**16.79**	37.01	12.68	**3.21**	27.63	7.06	**0.31**	19.32	5.29	**0.25**	8.73	3.96	**0.16**
	StyleGAN	99.98	100.00	**89.05**	96.96	79.88	**55.06**	87.52	44.37	**11.01**	71.06	22.59	**5.60**	35.83	11.44	**1.18**
	WebMorph	94.13	88.81	**21.04**	64.68	62.22	**6.94**	42.58	44.62	**0.57**	22.76	32.22	**0.33**	7.17	22.29	**0.16**
LMA-DRD	D	99.87	98.81	**93.18**	95.14	90.48	**75.42**	88.88	67.90	**57.39**	83.26	59.86	**40.87**	73.99	40.94	**17.10**
	PS	100.00	100.00	**99.22**	99.41	99.91	**92.15**	91.63	85.33	**66.59**	83.39	79.02	**52.74**	70.97	63.37	**37.07**
MIPGAN II	MIPGAN II	99.77	98.60	**97.45**	96.72	93.25	**80.01**	79.29	68.06	**43.47**	56.07	54.10	**34.60**	44.05	48.09	**17.53**
MorDiff	MorDiff	99.35	**94.31**	95.04	92.03	79.06	**69.31**	73.76	67.15	**47.84**	63.04	45.77	**29.73**	49.13	28.92	**14.08**
MorGAN	MorGAN	100.00	**99.25**	100.00	99.80	**97.80**	100.00	98.13	**89.16**	99.10	95.38	**83.63**	95.13	86.45	**74.34**	**83.22**
	LMA	100.00	99.85	**99.75**	99.71	99.50	**94.06**	99.10	95.52	**68.86**	95.76	89.99	**49.46**	85.67	79.99	**27.60**
Greedy	DIM	**85.64**	94.66	89.18	**26.13**	69.50	49.10	**7.36**	20.12	21.84	**5.87**	13.44	14.56	**2.60**	6.00	3.00
Average	–	94.60	85.17	**73.09**	79.68	71.97	**54.05**	66.58	55.86	**33.67**	56.88	46.8	**25.32**	43.88	36.64	**15.69**

ZS = Zero-Shot; *HO = Classification Head-Only baseline*; *LoRA = Low-Rank Adaptation fine-tuned*

without catastrophic forgetting, validates that parameter-efficient adaptation successfully specialized the model for morphing detection while preserving its foundational knowledge. The consistent improvements across diverse morphing techniques demonstrate that LoRA fine-tuning enhances rather than disrupts the model's pre-existing capabilities.

Beyond detection accuracy improvements, the fine-tuned model delivers substantial practical advantages in computational efficiency. The adapted *Gemma-3 12B* achieves a **30× inference speedup** compared to zero-shot evaluation, reducing processing time from 30 seconds to under 1 second per image. This dramatic optimization stems from three factors: the smaller model size (12B versus 27B parameters), simplified prompting that eliminates complex multi-step reasoning, and critically, the shift from generative text production to direct classification.

Table 5.19: **Bona Fide Presentation Classification Error Rate (BPCER)** at fixed **Attack Presentation Classification Error Rate (APCER)** thresholds for *Gemma-3* model variants. Comparison includes *Zero-Shot (27B)*, *Classification Head-Only baseline (12B frozen)*, and *LoRA fine-tuned (12B)* configurations across five operating points (0.1%, 1%, 5%, 10%, 20% APCER). Lower **BPCER values indicate fewer false rejections of genuine users** at each attack detection level. Best results per operating point are highlighted in **bold**.

Dataset	**Technique**	**@0.1%**			**@1%**			**@5%**			**@10%**			**@20%**		
		ZS	HO	LoRA	ZS	HO	LoRA	ZS	HO	LoRA	ZS	HO	LoRA	ZS	HO	LoRA
FERET	FaceMorpher	97.42	99.19	**83.53**	73.12	91.85	**47.59**	46.24	59.25	**19.58**	31.09	24.69	**8.06**	14.76	3.75	**1.84**
	OpenCV	96.84	97.29	**79.66**	85.19	72.92	**38.23**	58.83	33.51	**12.75**	42.24	13.48	**4.97**	20.10	3.96	**1.20**
	StyleGAN	99.85	99.42	**99.06**	98.61	94.21	**86.18**	93.13	**71.03**	71.23	84.98	**50.18**	61.56	66.72	**34.49**	48.21
FRGC	FaceMorpher	98.69	99.72	**74.76**	93.22	97.23	**21.89**	76.46	86.14	**4.91**	64.86	72.29	**2.15**	45.84	44.58	**0.98**
	OpenCV	99.35	99.80	**63.83**	96.45	97.96	**32.82**	88.53	89.81	**14.11**	81.86	79.62	**8.31**	70.99	59.24	**3.07**
	StyleGAN	99.90	99.87	**74.02**	98.95	98.70	**54.60**	94.82	93.51	**34.35**	90.98	87.01	**25.21**	83.61	74.03	**15.26**
FRLL	AMSL	100.00	99.90	**69.03**	83.23	98.96	**42.52**	73.24	94.81	**20.71**	69.26	89.62	**14.22**	44.28	79.24	**6.37**
	FaceMorpher	80.88	99.53	**31.38**	53.33	95.34	**0.49**	22.06	76.70	**0.00**	14.22	53.41	**0.00**	11.39	15.49	**0.00**
	OpenCV	89.28	97.95	**22.01**	69.20	79.52	**2.25**	32.34	14.17	**0.49**	19.39	0.98	**0.00**	9.66	0.49	**0.00**
	StyleGAN	96.68	99.15	**36.98**	86.77	91.46	**22.06**	68.85	57.29	**10.29**	51.53	22.55	**5.39**	37.48	13.24	**3.43**
	WebMorph	96.30	99.61	**36.51**	71.61	96.09	**3.82**	27.25	80.47	**1.47**	14.22	60.94	**0.49**	12.25	23.53	**0.49**
LMA	D	99.67	99.79	**95.75**	96.71	97.86	**72.49**	88.35	89.32	**48.63**	81.70	78.65	**33.02**	69.12	57.29	**17.80**
	PS	99.61	97.52	**96.97**	97.42	**81.57**	82.53	83.01	74.11	**63.59**	75.58	65.39	**46.25**	61.97	56.48	**35.33**
MIPGAN II	MIPGAN II	99.77	99.83	**62.18**	88.77	98.31	**55.00**	76.62	91.56	**35.86**	62.93	83.13	**23.57**	52.35	66.25	**15.29**
MorDiff	MorDiff	98.87	99.58	**63.81**	98.36	95.84	**48.68**	82.45	79.19	**32.08**	66.42	58.39	**26.04**	50.48	35.09	**15.85**
MorGAN	MorGAN	**99.26**	99.89	99.84	**97.77**	98.92	98.36	94.75	94.62	**88.81**	89.85	89.23	**82.29**	84.07	78.46	**69.05**
	LMA	99.93	99.90	**89.29**	98.39	99.00	**74.56**	90.96	95.00	**55.09**	87.29	90.00	**39.69**	80.45	80.00	**24.72**
Greedy	DIM	94.61	80.39	**71.57**	**38.24**	47.06	41.18	**12.99**	22.55	13.73	**2.45**	11.76	11.76	**1.47**	4.90	6.37
Average	–	97.05	97.69	**69.45**	84.74	90.71	**45.85**	67.27	72.39	**29.32**	57.27	57.30	**21.83**	45.39	40.58	**14.74**

ZS = Zero-Shot; *HO = Classification Head-Only baseline*; *LoRA = Low-Rank Adaptation fine-tuned*

Performance by Morphing Technique Category. Analysis by morphing generation method reveals distinct patterns. For landmark-based morphs (*OpenCV*, *FaceMorpher*, *WebMorph*, *AMSL*, *LMA*), LoRA fine-tuning achieved exceptional results. *FRLL-FaceMorpher* dropped from 18.15% (baseline) to 0.49% EER, *FRLL-OpenCV* from 6.38% to 1.47%, and *FRLL-WebMorph* from 21.47% to 2.37%. These dramatic improvements indicate that fine-tuning enables the model to reliably detect the geometric inconsistencies and blending artifacts characteristic of landmark-based methods.

GAN-based morphs (*StyleGAN*, *MIPGAN-II*, *MorGAN*) showed more variable responses to fine-tuning. While *FRGC-StyleGAN* improved substantially from 40.82% to 17.32% EER, and *MIPGAN-II* from 31.67% to 17.92%, the *MorGAN* dataset remained challenging with only modest gains (46.69% to 45.26%). This suggests that certain GAN morphs that closely approximate genuine facial

distributions remain inherently difficult to detect even with domain adaptation.

Diffusion-based morphs (*Greedy-DiM*, *MorDiff*) exhibited mixed results. *MorDiff* benefited from fine-tuning (24.88% to 17.33% EER), while *Greedy-DiM* showed slight degradation (11.24% to 11.78%). Notably, the frozen baseline performed surprisingly well on *Greedy-DiM*, suggesting the pre-trained model already possesses knowledge relevant to detecting diffusion artifacts.

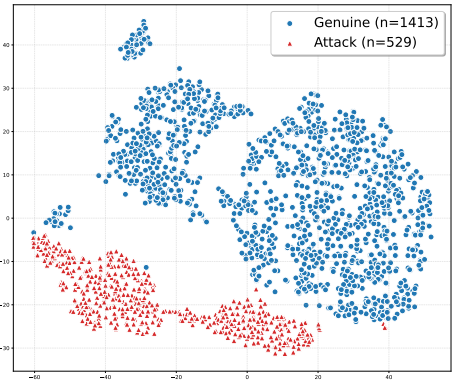
Operating Point Analysis. Tables 5.18 and 5.19 reveal the practical implications of these improvements. At a security-critical 1% BPCER threshold, the frozen baseline achieved 71.97% average APCER, while LoRA reduced this to 54.05%. This 17.92 percentage point improvement.

The biggest improvements occur for landmark-based morphs at moderate security levels. At 5% BPCER, *FRLL-FaceMorpher* APCER decreased from 25.99% (baseline) to 0.16% (LoRA), effectively achieving near-perfect detection. Similar patterns appear across *FRLL-OpenCV* (7.06% $\rightarrow$ 0.31%) and *FRGC-FaceMorpher* (39.80% $\rightarrow$ 4.98%), demonstrating that fine-tuning particularly enhances detection of traditional morphing artifacts.

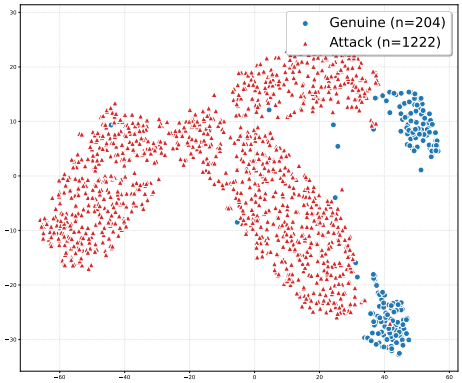
The reciprocal analysis (Table 5.19) confirms operational improvements. To achieve 95% attack detection (5% APCER), the frozen baseline would reject 72.39% of legitimate users on average, while LoRA reduces this to 29.32%. At balanced operating points (10% APCER), LoRA achieves 21.83% average BPCER compared to 57.30% for the baseline.

Feature Space Analysis . To validate the quantitative results, we examine t-SNE visualizations of the learned feature representations, which reveal the model's ability to separate genuine and morphed samples in the high-dimensional embedding space.

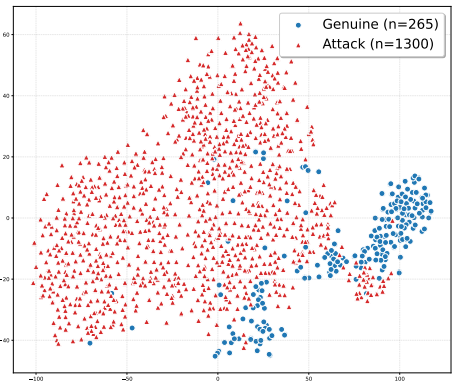
StyleGAN Morphs: The comparison between *FERET-StyleGAN* (34.97% EER) and *FRLL-StyleGAN* (6.83% EER) presents a performance-feature mismatch. Figure 5.2a shows that *FERET-StyleGAN* exhibits excellent cluster separation with distinct boundaries between genuine (blue) and attack (red) samples,



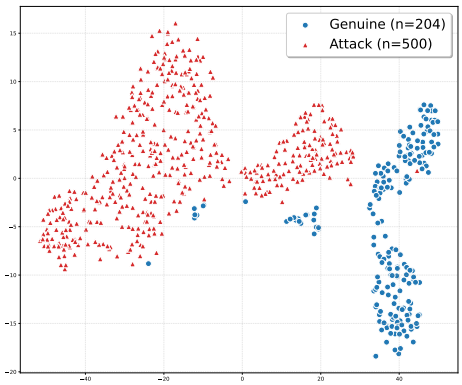
(a) FERET-StyleGAN



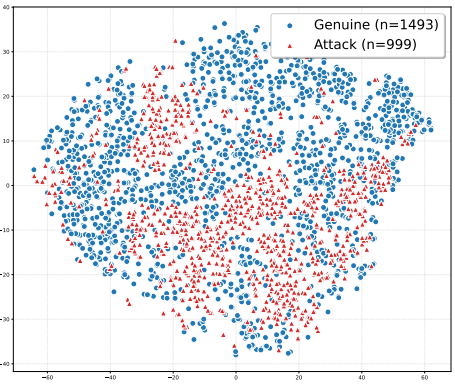
(b) FRLL-StyleGAN



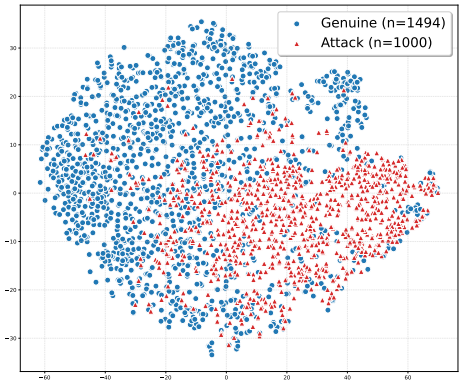
(c) MorDiff



(d) Greedy-DiM



(e) MorGAN-GAN



(f) MorGAN-LMA

Figure 5.2: t-SNE visualization of feature representations for different morphing techniques. Each plot shows the two-dimensional projection of high-dimensional feature embeddings, where axes represent t-SNE Component 1 (horizontal) and t-SNE Component 2 (vertical). Blue points indicate genuine samples, red points indicate attack samples. The degree of cluster separation demonstrates the model's ability to distinguish between genuine and morphed images for each attack type.

while *FRLL-StyleGAN* in Figure 5.2b, despite superior EER performance, shows comparable but not dramatically better separability. This discrepancy suggests a classification head calibration issue where the learned features achieve good separation, but the final classification layer fails to properly leverage this separability for the *FERET* dataset. The classification head, may be poorly calibrated for the specific feature distributions of *FERET* images, leading to suboptimal decision boundaries despite good representational quality.

Diffusion-Based Morphs: Both *MorDiff* (17.33% EER) in Figure 5.2c and *Greedy-DiM* (11.78% EER) in Figure 5.2d exhibit a characteristic pattern where genuine samples form tight, well-defined clusters while attack samples are more dispersed throughout the feature space. This scattered distribution of diffusion-generated morphs reflects their sophisticated nature, these attacks span a broader range of the learned manifold, making classification more challenging. The relatively moderate EER values, despite this dispersion in attack samples, demonstrate the model's ability to maintain reliable genuine sample clustering, which provides a stable reference for detection decisions.

Challenging Datasets: *MorGAN-GAN* (45.26% EER) in Figure 5.2e and *MorGAN-LMA* (23.% EER) in Figure 5.2f show fundamentally different feature distributions, with both genuine and attack samples scattered across the embedding space in patterns resembling the untrained representations from Figure 4.6a. This contrasts sharply with other morphing techniques: *StyleGAN* variants demonstrate clear GAN-type clustering with coherent attack distributions, while genuine samples across all other datasets (*FERET*, *FRLL*, *MorDiff*, *Greedy-DiM*) form tight, well-defined clusters regardless of the attack type.

The poor separability in both *MorGAN-GAN* (GAN-based) and *MorGAN-LMA* (landmark-based) datasets is particularly notable given that other GAN attacks (*StyleGAN*) and landmark attacks show distinct clustering patterns. This suggests the issue transcends morphing methodology. Both datasets use 64×64 pixel images, significantly smaller than the model's expected input resolution. This resolution mismatch prevents extraction of discriminative features, as fine-grained artifacts necessary for morphing detection are lost during downsampling. The scattered *bona fide* representations, unlike the tight genuine clusters in all other datasets, indicate the model cannot establish reliable reference points for

these low-resolution inputs, explaining the consistently high error rates across both attack types.

Key Findings. The evaluation establishes four critical insights.

First, the frozen *Gemma-3 12B* model demonstrates **substantial inherent capability for morphing detection**, validating that foundation models possess relevant forensic knowledge without task-specific training.

Second, LoRA fine-tuning consistently enhances this capability, **with average improvements of** 12.52 **percentage points** over the baseline.

Third, the effectiveness of adaptation varies by morphing technique, with **landmark-based methods showing exceptional responsiveness** while sophisticated GAN morphs remain challenging.

These results demonstrate that parameter-efficient fine-tuning successfully transforms a general-purpose multimodal LLM into a competitive morphing attack detector, with performance approaching specialized systems for certain attack types while maintaining the model's foundational capabilities.

Fourth, visual feature analysis reveals that input resolution critically impacts detection capability regardless of morphing technique, both GAN-based (*MorGAN*) and landmark-based (*LMA-PS*) attacks at 64×64 resolution show scattered feature distributions and poor genuine sample clustering, contrasting with tight clustering patterns observed across all higher-resolution datasets, demonstrating that there is a lower limit to input resolution for effective morphing detection.

5.3 Comparative Analysis with SOTA Models and MAD Baselines

In this section, we evaluate the performance of our fine-tuned *Gemma-3* model against established morphing attack detection (MAD) methods to contextualize

its capabilities within the current state-of-the-art. We compare against three categories of baselines: supervised deep learning methods (*MixFaceNet-MAD*, *PW-MAD*, *Inception-MAD*), unsupervised approaches (*SPL-MAD*, *MAD-DDPM*), and other foundation model adaptations for MAD (*MADation*).

This comprehensive benchmarking across diverse architectural paradigms and training methodologies provides critical insights into the relative strengths and limitations of multimodal LLM-based detection compared to both purpose-built MAD systems and alternative foundation model approaches.

5.3.1 Supervised Deep Learning Baseline Comparison

In this section, we evaluate our fine-tuned *Gemma-3 12B-MAD* model against established supervised deep learning baselines including *MixFaceNet-MAD*, *PW-MAD*, and *Inception-MAD*. We assess cross-dataset generalization by comparing performance across multiple training and evaluation configurations following the *SPL-MAD* framework [6].

Table 5.20: Equal Error Rate (EER, %) comparison of supervised MAD baselines trained on different datasets (columns) and evaluated on multiple test sets (rows), following the structure of SPL-MAD Table 2. Asterisks (*) denote intra-dataset evaluation as in the source. The rightmost column reports our *Gemma3-12B-MAD* (LoRA) results for the corresponding test sets.

Test data		***MixFaceNet-MAD [21]***					***PW-MAD [5]***					***Inception-MAD [5]***					***Gemma3-12B-MAD***
		D	**PS**	**-LMA**	**-GAN**	**SMDD**	**D**	**PS**	**-LMA**	**-GAN**	**SMDD**	**D**	**PS**	**-LMA**	**-GAN**	**SMDD**	**EER**
LMAD-DRD	*D*	15.68*	18.03	17.06	25.01	19.42	20.80*	25.10	22.34	40.21	17.06	7.64*	17.06	15.68	50.77	**15.11**	18.90
	PS	21.77	18.44*	27.05	27.05	23.72	26.48	23.72*	29.41	44.11	20.39	**11.37**	12.75*	22.34	38.42	19.01	28.28
MorGAN	*LMA*	39.42	**22.89**	10.61*	46.42	30.12	34.20	34.14	9.71*	34.37	27.31	38.55	31.73	8.43*	40.16	28.51	23.06
	GAN	53.01	50.44	42.57	24.90*	42.64	52.04	46.59	42.80	8.84*	43.78	50.84	38.79	**27.41**	0.40*	44.34	45.26
FRLL-Morphs	*OpenCV*	8.82	13.22	8.91	17.66	4.39	17.33	15.69	13.96	45.59	2.42	13.72	10.76	6.86	55.89	5.38	**1.47**
	FaceMorpher	7.80	10.97	7.34	15.65	3.87	13.88	15.14	10.92	44.57	2.20	16.62	15.81	6.32	66.14	3.17	**0.49**
	StyleGAN2	20.07	15.29	13.41	23.51	8.89	29.97	27.64	18.11	48.53	16.64	37.24	19.58	20.56	55.03	11.37	**6.83**
	WebMorph	25.97	29.04	20.61	30.39	12.35	33.78	28.51	35.75	52.43	16.65	57.38	58.32	30.88	77.42	9.86	**2.37**
	AMSL	24.53	27.59	19.24	30.03	15.18	36.25	32.95	34.38	48.52	15.18	49.02	61.44	**9.80**	86.49	10.79	12.74
FERET-Morphs	*OpenCV*	28.12	32.19	31.57	33.86	31.74	37.27	45.29	34.27	43.11	39.93	**6.39**	7.23	42.12	13.62	59.32	7.40
	FaceMorpher	22.57	29.48	27.90	31.81	23.69	35.16	44.30	28.24	40.40	29.41	**5.17**	6.91	36.53	18.36	46.94	9.30
	StyleGAN2	29.57	29.02	35.46	39.41	39.85	44.25	45.30	29.70	42.47	47.20	9.03	**7.12**	35.29	15.09	60.05	34.97
FRGC-Morphs	*OpenCV*	23.81	25.04	31.62	21.11	20.67	57.06	48.60	29.74	53.55	26.45	34.32	13.65	36.17	59.66	19.63	**9.23**
	FaceMorpher	22.83	23.54	29.38	19.98	18.10	56.00	50.70	30.49	51.61	23.40	34.96	19.71	35.10	56.91	16.06	**4.98**
	StyleGAN2	32.71	28.68	21.70	21.95	**11.62**	37.38	38.42	16.43	26.62	14.32	41.14	25.85	36.19	47.03	15.26	17.32
Greedy-DiM	*Diffusion*	45.10	41.67	40.69	48.04	39.71	17.16	33.82	17.16	15.20	42.16	31.86	51.96	25.98	29.90	56.86	**11.78**
MorDiff	*Diffusion*	21.30	23.70	28.83	30.19	20.40	3.21	**0.98**	11.60	16.00	13.80	21.08	21.78	19.41	56.09	15.23	17.33
Average performance†		26.71	26.30	25.21	28.88	21.55	33.21	33.32	25.33	40.46	23.43	28.67	25.48	25.42	47.94	25.70	**14.81**

*Intra-dataset evaluation as reported in the source. †Averages for baselines from SPL-MAD Table 2 [6] (excluding intra-dataset). Greedy-DiM baseline values from SelfMAD Table [1]. MorDiff baseline values from our implementations.

Table 5.21: BPCER (%) at fixed APCER = 5% and 10% for supervised MAD baselines trained on **LMA-DRD-D (Digital)** dataset and evaluated on multiple test sets. Lower is better.

Test Data		MixFaceNet-MAD [21]		PW-MAD [5]		Inception-MAD [5]		Gemma3-12B-MAD	
		5%	10%	5%	10%	5%	10%	5%	10%
FRGC	*FaceMorpher*	85.89	73.44	95.12	86.00	96.89	90.15	**4.91**	**2.15**
	OpenCV	51.97	37.03	91.29	83.51	93.57	80.29	**14.11**	**8.31**
	StyleGAN2	64.63	41.91	97.82	93.67	89.00	78.11	**34.35**	**25.21**
FERET	*FaceMorpher*	69.38	57.09	47.26	25.33	31.57	21.55	**19.58**	**8.06**
	OpenCV	79.40	68.43	21.55	13.99	32.33	20.98	**12.75**	**4.97**
	StyleGAN2	91.87	81.10	95.84	88.09	**21.93**	**13.04**	71.23	61.56
FRLL	*AMSL*	71.40	61.47	87.08	71.95	99.26	98.48	**20.71**	**14.22**
	FaceMorpher	23.71	13.66	69.85	33.51	81.36	70.45	**0.00**	**0.00**
	OpenCV	39.97	18.02	22.19	12.04	69.45	57.82	**0.49**	**0.00**
	StyleGAN2	54.58	39.20	93.94	80.61	67.59	57.77	**10.29**	**5.39**
	WebMorph	85.26	71.01	90.09	75.68	99.92	99.75	**1.47**	**0.49**
Greedy-DiM	*Diffusion*	93.40	85.60	64.20	36.60	85.40	77.40	**13.73**	**11.76**
MorDiff	*Diffusion*	50.98	42.16	**1.96**	**0.49**	50.98	37.74	32.08	26.04
Average		66.34	53.09	67.55	53.96	70.71	61.81	**18.13**	**12.94**

Table 5.20 presents a comprehensive comparison between our fine-tuned *Gemma-3 12B-MAD* model and three established supervised deep learning baselines following the evaluation framework of *SPL-MAD* [6]. Each baseline method was trained independently on five distinct datasets: LMA-DRD Digital (D), LMA-DRD Print-Scan (PS), MorGAN-LMA, MorGAN-GAN, and SMDD, resulting in 15 model variants evaluated across diverse morphing attack types.

Overall Performance Comparison. The *Gemma-3 12B-MAD* model achieves an average Equal Error Rate (EER) of 14.81%, substantially outperforming all supervised baseline configurations. Among the traditional methods, models trained on SMDD demonstrated the best average performance across architectures: *MixFaceNet-MAD* (21.55%), *PW-MAD* (23.48%), and *Inception-MAD* (25.70%). This 6.74 percentage point improvement over the best baseline average underscores the advantage of combining foundation model knowledge with domain-specific fine-tuning.

The performance gap widens dramatically at operational thresholds. Table 5.25 reveals that at a security-critical 5% APCER, *Gemma-3* achieves an

Table 5.22: BPCER (%) at fixed APCER = 5% and 10% for supervised MAD baselines trained on **LMA-DRD-PS (Print-Scan)** dataset and evaluated on multiple test sets. Lower is better.

Test Data		MixFaceNet-MAD [21]		PW-MAD [5]		Inception-MAD [5]		Gemma3-12B-MAD	
		5%	10%	5%	10%	5%	10%	5%	10%
FRGC	*FaceMorpher*	75.31	55.19	94.92	89.52	93.36	85.89	**4.91**	**2.15**
	OpenCV	79.25	64.83	98.86	97.20	53.11	37.34	**14.11**	**8.31**
	StyleGAN2	46.89	33.82	66.29	59.13	47.10	38.28	**34.35**	**25.21**
FERET	*FaceMorpher*	82.61	73.53	52.93	32.70	69.57	57.84	**19.58**	**8.06**
	OpenCV	77.13	66.16	17.39	12.29	64.84	53.69	**12.75**	**4.97**
	StyleGAN2	83.93	68.43	97.16	93.95	**31.38**	**22.12**	71.23	61.56
FRLL	*AMSL*	56.55	45.06	53.75	46.85	99.49	98.16	**20.71**	**14.22**
	FaceMorpher	15.38	8.33	34.28	22.42	78.95	68.99	**0.00**	**0.00**
	OpenCV	28.26	14.00	12.29	6.72	92.79	88.94	**0.49**	**0.00**
	StyleGAN2	28.07	17.76	78.15	67.02	75.04	68.33	**10.29**	**5.39**
	WebMorph	76.82	65.36	66.67	48.81	93.28	87.96	**1.47**	**0.49**
Greedy-DiM	*Diffusion*	89.00	78.80	83.40	70.00	95.20	92.20	**13.73**	**11.76**
MorDiff	*Diffusion*	57.28	42.46	**0.49**	**0.49**	53.92	41.18	32.08	26.04
Average		61.27	48.75	58.20	49.78	72.93	64.69	**18.13**	**12.94**

average BPCER of 18.13%, compared to 39.85% for the best-performing *PW-MAD* trained on SMDD. This represents a 54.5% reduction in false rejections while maintaining equivalent attack detection rates, transforming the system from marginally viable to operationally practical.

Performance by Morphing Technique. Analysis by morphing generation method reveals distinct patterns in model capabilities. For landmark-based morphs (*OpenCV*, *FaceMorpher*, *WebMorph*, *AMSL*), *Gemma-3* demonstrates exceptional performance, achieving near-perfect detection on FRLL datasets: OpenCV (1.47%), FaceMorpher (0.49%), and WebMorph (2.37%). In contrast, the best supervised baseline (*MixFaceNet-MAD* trained on SMDD) achieves 4.39%,3.87%, and 12.35% respectively on these same attacks. This superiority extends to operational metrics, where *Gemma-3* maintains near-zero BPCER at both 5% and 10% APCER for FRLL landmark morphs.

GAN-based morphs present a more nuanced picture. While *Gemma-3* achieves competitive results on FRLL-StyleGAN2 (6.83% vs. 8.89% for *MixFaceNet-SMDD*), it struggles with MorGAN-GAN attacks (45.26%), perform-

Table 5.23: BPCER (%) at fixed APCER = 5% and 10% for supervised MAD baselines trained on **MorGAN–LMA** dataset and evaluated on multiple test sets. Lower is better.

Test Data		MixFaceNet-MAD [21]		PW-MAD [5]		Inception-MAD [5]		Gemma3-12B-MAD	
		5%	10%	5%	10%	5%	10%	5%	10%
FRGC	*FaceMorpher*	85.58	76.35	93.67	66.29	79.46	60.48	**4.91**	**2.15**
	OpenCV	85.06	72.93	48.86	20.12	89.00	76.04	**14.11**	**8.31**
	StyleGAN2	49.38	30.81	61.00	39.21	91.70	84.34	**34.35**	**25.21**
FERET	*FaceMorpher*	84.12	76.75	75.99	48.58	70.32	57.47	**19.58**	**8.06**
	OpenCV	89.41	78.45	75.99	65.97	86.77	73.72	**12.75**	**4.97**
	StyleGAN2	76.37	61.63	77.69	**60.87**	81.85	68.81	**71.23**	61.56
FRLL	*AMSL*	84.64	74.53	47.49	33.47	35.26	20.37	**20.71**	**14.22**
	FaceMorpher	69.93	49.05	26.72	12.54	20.10	5.76	**0.00**	**0.00**
	OpenCV	44.64	27.27	27.35	19.66	63.47	41.52	**0.49**	**0.00**
	StyleGAN2	83.39	71.11	42.39	30.20	66.94	48.61	**10.29**	**5.39**
	WebMorph	94.35	90.01	92.22	83.70	72.73	64.13	**1.47**	**0.49**
Greedy-DiM	*Diffusion*	93.40	88.00	40.80	28.60	66.00	46.00	**13.73**	**11.76**
MorDiff	*Diffusion*	51.96	45.56	**27.45**	**17.65**	55.39	37.75	32.08	26.04
Average		76.32	64.81	56.74	40.53	67.61	52.69	**18.13**	**12.94**

ing comparably to baseline methods. However, even on these challenging morphs, *Gemma-3* maintains better operational characteristics, with substantially lower BPCER at fixed APCER thresholds.

Diffusion-based morphs reveal interesting generalization patterns. On Greedy-DiM, *Gemma-3* (11.78%) outperforms most baseline configurations, with only *PW-MAD* trained on MorGAN-LMA achieving comparable results (17.16%). For MorDiff, *PW-MAD* trained on D and PS datasets achieves exceptional performance (0.98% EER), suggesting that specific training data distributions can optimize detection for particular diffusion techniques.

Cross-Dataset Generalization. The supervised baselines exhibit significant performance degradation in cross-dataset evaluation. Models trained on LMA-DRD Digital achieve strong intra-dataset performance (7.64% for *Inception-MAD*) but struggle when evaluated on morphs from different sources, with EERs exceeding 50% on StyleGAN morphs from FRLL and FRGC datasets. This pattern repeats across all baseline architectures and training configurations, highlighting the overfitting endemic to supervised MAD approaches.

Table 5.24: BPCER (%) at fixed APCER = 5% and 10% for supervised MAD baselines trained on **MorGAN–GAN** dataset and evaluated on multiple test sets. Lower is better.

Test Data		**MixFaceNet-MAD [21]**		**PW-MAD [5]**		**Inception-MAD [5]**		**Gemma3-12B-MAD**	
		5%	**10%**	**5%**	**10%**	**5%**	**10%**	**5%**	**10%**
FRGC	*FaceMorpher*	76.45	60.89	99.79	98.76	99.69	98.65	**4.91**	**2.15**
	OpenCV	80.71	68.78	99.07	97.72	98.76	96.58	**14.11**	**8.31**
	StyleGAN2	89.94	79.36	100.00	100.00	95.23	89.83	**34.35**	**25.21**
FERET	*FaceMorpher*	95.46	90.74	31.95	16.82	95.65	92.25	**19.58**	**8.06**
	OpenCV	90.55	83.55	70.51	54.63	98.11	96.98	**12.75**	**4.97**
	StyleGAN2	94.90	88.66	88.66	80.15	82.99	72.40	**71.23**	**61.56**
FRLL	*AMSL*	95.45	88.83	92.41	88.00	100.00	100.00	**20.71**	**14.22**
	FaceMorpher	97.08	92.35	66.07	48.80	99.31	98.71	**0.00**	**0.00**
	OpenCV	96.56	93.94	97.22	83.21	98.94	96.97	**0.49**	**0.00**
	StyleGAN2	85.76	78.97	97.46	90.59	98.53	97.55	**10.29**	**5.39**
	WebMorph	95.25	92.55	93.61	89.76	99.84	99.43	**1.47**	**0.49**
Greedy-DiM	*Diffusion*	97.00	93.40	39.60	22.20	86.60	73.00	**13.73**	**11.76**
MorDiff	*Diffusion*	51.96	44.12	39.71	27.10	96.95	93.89	**32.08**	**26.04**
Average		88.23	81.24	78.16	69.06	96.02	92.79	**18.13**	**12.94**

In contrast, *Gemma-3*'s consistent performance across diverse test sets demonstrates superior generalization. The model maintains reliable detection across different imaging conditions (FRLL vs. FRGC vs. FERET), morphing techniques, and post-processing scenarios (digital vs. print-scan), suggesting that foundation model pre-training provides robust feature representations that transcend dataset-specific artifacts.

Key Findings. The evaluation establishes three critical insights.

First, LLM fine-tuning achieves **superior average performance (**14.81% **EER) compared to purpose-built supervised detectors (**21.55% **best average)**, with the advantage most pronounced at operational thresholds.

Second, *Gemma-3* demonstrates **exceptional capability on landmark-based morphs** while maintaining competitive performance on GAN and diffusion attacks, whereas supervised baselines typically excel in narrow domains corresponding to their training data.

Table 5.25: BPCER (%) at fixed APCER = 5% and 10% for supervised MAD baselines trained on **SMDD** dataset and evaluated on multiple test sets. Lower is better.

Test Data		MixFaceNet-MAD [21]		PW-MAD [5]		Inception-MAD [5]		Gemma3-12B-MAD	
		5%	10%	5%	10%	5%	10%	5%	10%
FRGC	*FaceMorpher*	33.40	20.95	51.76	29.25	54.15	35.68	**4.91**	**2.15**
	OpenCV	35.27	20.95	56.85	38.90	67.12	47.61	**14.11**	**8.31**
	StyleGAN2	**26.97**	**15.87**	35.68	21.06	99.90	97.20	**34.35**	**25.21**
FERET	*FaceMorpher*	53.69	34.97	42.34	**5.10**	69.57	46.12	**19.58**	8.06
	OpenCV	53.88	39.51	74.48	35.92	100.00	99.43	**12.75**	**4.97**
	StyleGAN2	87.33	75.24	81.47	**53.31**	99.43	93.76	**71.23**	61.56
FRLL	*AMSL*	65.56	60.00	**4.64**	**2.21**	11.03	7.17	20.71	14.22
	FaceMorpher	11.51	8.16	1.72	1.29	2.66	2.23	**0.00**	**0.00**
	OpenCV	15.89	10.07	1.80	0.90	37.59	22.60	**0.49**	**0.00**
	StyleGAN2	84.94	76.27	34.62	26.43	45.58	35.84	**10.29**	**5.39**
	WebMorph	78.38	67.16	12.29	9.75	18.18	12.69	**1.47**	**0.49**
Greedy-DiM	*Diffusion*	90.20	86.20	95.40	86.80	100.00	99.80	**13.73**	**11.76**
MorDiff	*Diffusion*	68.87	40.69	**25.00**	**20.59**	42.65	23.04	32.08	26.04
Average		54.30	42.77	39.85	25.50	57.53	47.94	**18.13**	**12.94**

Third, the substantial reduction in BPCER at fixed APCER thresholds **(**18.13% **vs.** 39.85% **at** 5% **APCER) represents a** 54.5% **decrease in false rejections**, demonstrating that LLM adaptation achieves superior operational metrics that fundamentally improve the balance between detection accuracy and user convenience compared to traditional supervised approaches.

5.3.2 Unsupervised and Self-Supervised Baseline Comparison

In this section, we evaluate *Gemma-3 12B-MAD* against state-of-the-art unsupervised and self-supervised morphing attack detection methods including *SelfMAD*, *SPL-MAD*, *MAD-DDPM*, and quality-based approaches. We assess performance across diverse morphing techniques to establish our method's position within the current unsupervised detection paradigm.

Table 5.26 presents a comprehensive evaluation comparing *Gemma-3 12B-MAD* against state-of-the-art unsupervised and self-supervised morphing attack detection methods following the protocol established by *SelfMAD*. The comparison includes quality-based approaches (*FIQA-MagFace*, *CNNIQA*), self-

Table 5.26: Equal Error Rate (EER %) comparison of Gemma3-12B-MAD with SOTA unsupervised models using the protocol from SelfMAD [1]. Lower is better.

Test Data		FIQA-MagFace [51]	CNNIQA [51]	SPL-MAD [6]	MAD-DDPM [40]	SBI [88]	SelfMAD [1]	Gemma3-12B-MAD
FRGC-M	FM	33.82	42.84	16.91	25.62	16.68	5.59	**4.98**
	OCV	33.30	43.15	20.75	28.22	15.32	**2.59**	9.23
	SG	**14.21**	36.51	16.80	9.02	52.90	15.84	17.32
FERET-M	FM	25.14	13.23	20.42	27.98	26.47	**3.19**	9.30
	OCV	26.14	20.45	25.71	31.38	28.73	**1.13**	7.40
	SG	**12.67**	33.84	25.33	32.14	41.83	18.14	34.97
FRLL-M	AMSL	30.94	21.61	3.26	27.13	11.76	**0.99**	12.74
	FM	27.99	19.97	1.03	10.40	13.73	**0.00**	0.49
	OCV	24.73	7.53	1.88	13.76	12.25	**0.00**	1.47
	SG	7.53	35.92	14.65	14.32	44.61	10.34	**6.83**
	WM	27.19	21.54	6.39	30.30	39.22	3.45	**2.37**
Greedy	DiM	47.00	49.40	37.72	36.10	33.82	**7.60**	11.78
Average		25.89	28.83	15.90	23.86	28.11	**5.74**	9.91

Results taken from SelfMAD paper [1]. *FM: FaceMorpher, OCV: OpenCV, SG: StyleGAN2, WM: Webmorph, BE@AE: BPCER@APCER

supervised methods (*SPL-MAD*, *MAD-DDPM*, *SBI*), and the current state-of-the-art self-supervised approach *SelfMAD*.

Overall Performance Comparison. *SelfMAD* establishes the current state-of-the-art with an average Equal Error Rate (EER) of 5.74%, demonstrating the effectiveness of self-supervised learning with synthetic morphing artifacts. *Gemma-3 12B-MAD* achieves an average EER of 9.91%, placing it as the second-best performer, substantially outperforming now traditional unsupervised methods including *SPL-MAD* (15.90%), *FIQA-MagFace* (25.89%), *MAD-DDPM* (23.86%), and *CNNIQA* (28.83%). While not surpassing *SelfMAD* in average EER, *Gemma-3 12B-MAD* exhibits competitive performance at operational thresholds, achieving comparable BPCER values at fixed APCER points.

Performance by Morphing Technique Category. Analysis reveals distinct performance patterns across morphing generation methods. For landmark-based morphs, *SelfMAD* demonstrates exceptional capability, achieving near-perfect detection on FRLL datasets: OpenCV (0.00% EER), FaceMorpher (0.00% EER), and remarkably low rates on AMSL (0.99%) and WebMorph (3.45%). *Gemma-3 12B-MAD*, while not matching these exceptional results, still achieves strong per-

Table 5.27: BPCER (%) at fixed APCER = 5% and 10% for Gemma3-12B-MAD compared with SOTA unsupervised models using the protocol from SelfMAD [1]. Lower is better.

Test Data		FIQA-MagFace [51]		CNNIQA [51]		SPL-MAD [6]		MAD-DDPM [40]		SBI [88]		SelfMAD [1]		Gemma3-12B-MAD	
		5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%
FRGC-M	FM	73.79	62.84	75.94	66.86	25.39	21.47	95.12	90.15	38.07	26.14	6.43	2.80	**4.91**	**2.15**
	OCV	74.71	62.52	74.64	66.35	32.50	25.42	95.12	90.15	36.31	25.10	**1.14**	**0.41**	14.11	8.31
	SG	26.46	17.60	70.34	57.93	**26.13**	**21.09**	95.12	90.15	97.10	94.40	45.23	25.52	34.35	25.21
FERET-M	FM	61.22	44.44	35.17	19.32	40.85	27.09	95.27	90.17	60.87	52.36	**1.70**	**0.38**	19.50	8.04
	OCV	61.50	43.95	58.60	37.23	57.45	45.60	95.27	90.17	70.08	60.61	**0.57**	**0.38**	12.75	4.97
	SG	**24.63**	**15.71**	79.55	66.17	62.06	49.72	95.27	90.17	90.55	82.42	46.12	32.33	71.23	61.56
FRLL-M	AMSL	77.94	66.18	60.29	39.22	0.50	0.50	94.94	90.02	24.23	16.78	**0.05**	**0.05**	20.71	14.22
	FM	73.04	57.35	57.84	36.76	0.99	0.99	95.19	90.38	36.99	26.10	0.26	0.17	**0.00**	**0.00**
	OCV	66.18	53.43	11.76	4.41	0.50	0.50	95.17	90.01	27.85	18.84	**0.00**	**0.00**	0.49	**0.00**
	SG	8.82	5.39	75.49	68.14	32.18	24.75	95.17	90.18	94.68	90.92	24.22	12.52	**10.29**	**5.39**
	WM	68.14	55.39	46.57	33.33	11.39	3.47	95.09	90.34	89.93	83.37	1.64	0.41	**1.47**	**0.49**
Greedy	DiM	94.61	85.78	96.08	93.14	80.69	71.78	95.20	89.70	90.60	81.60	37.60	27.80	**13.73**	**11.76**
Average		59.25	47.55	61.86	49.07	30.89	24.36	95.16	90.13	63.11	54.89	**13.75**	**8.56**	16.96	11.84

Results taken from SelfMAD paper [1]. *FM: FaceMorpher, OCV: OpenCV, SG: StyleGAN2, WM: Webmorph, BE@AE: BPCER@APCER

formance with 1.47%, 0.49%, 12.74%, and 2.37% EER respectively, significantly outperforming all other baselines.

On GAN-based morphs (StyleGAN2), *Gemma-3 12B-MAD* demonstrates competitive or superior performance compared to *SelfMAD*. For FRLL-StyleGAN2, *Gemma-3 12B-MAD* achieves 6.83% EER versus *SelfMAD*'s 10.34%, and maintains better operational metrics with 10.29% BPCER at 5% APCER compared to *SelfMAD*'s 24.22%. This pattern suggests that foundation model pre-training provides robust features for detecting certain sophisticated GAN artifacts that challenge even specialized self-supervised approaches.

The most significant performance differential appears on diffusion-based morphs (Greedy-DiM). *SelfMAD* achieves 7.60% EER, demonstrating its strong generalization to novel morphing techniques. *Gemma-3 12B-MAD* records 11.78% EER but excels at operational thresholds, achieving 13.73% BPCER at 5% APCER compared to *SelfMAD*'s 37.60%, indicating better calibration for practical deployment despite higher overall error rates.

Comparison with Traditional Unsupervised Methods. Quality-based approaches (*FIQA-MagFace*, *CNNIQA*) show inconsistent performance across mor-

phing types. While *CNNIQA* achieves reasonable detection on certain landmark morphs (7.53% EER on FRLL-OpenCV), both methods fail considerably on others, with average EERs exceeding 25%. Their operational metrics are particularly poor, with BPCER values often exceeding 60% at 5% APCER, rendering them unsuitable for practical deployment.

SPL-MAD, despite being an early self-supervised approach, demonstrates competitive performance on FRLL landmark morphs (1–7% EER) but struggles with GAN and diffusion attacks. *MAD-DDPM* shows more consistent but moderate performance across all morphing types, with EERs ranging from 9–37%.

Operational Threshold Analysis. At security-critical operating points, *Gemma-3 12B-MAD* demonstrates strong practical viability. At 5% APCER, it achieves an average BPCER of 16.96%, compared to *SelfMAD*'s 13.75%. This 3.21 percentage point difference is smaller than the 4.17 point gap in average EER, indicating that *Gemma-3 12B-MAD* provides more suitable scores for threshold-based deployment.

Key Findings. The evaluation establishes that while ***SelfMAD* remains the state-of-the-art with its** 5.74% **average EER**, *Gemma-3 12B-MAD* represents a compelling alternative approach that achieves **the second-best overall performance (**9.91% **EER)** through foundation model adaptation. The comparison reveals complementary strengths: *SelfMAD* excels at landmark-based morphs, while in comparison *Gemma-3 12B-MAD* demonstrates superior or competitive performance on GAN-based attacks and maintains better operational characteristics on certain challenging datasets. Both approaches **substantially outperform traditional unsupervised methods, with average performance improvements exceeding 5–15 percentage points**, establishing a new performance tier for morphing attack detection without explicit supervision on real morphing attacks.

5.3.3 Gemma-3 12B-MAD vs. MADation and GPT4-Turbo

In this section, we evaluate *Gemma-3 12B-MAD* against *MADation* foundation model variants and *GPT-4 Turbo* to compare different foundation model adaptation strategies. We assess performance across FRLL-based morphing techniques and advanced GAN/diffusion datasets to establish our approach's position among state-of-the-art foundation model-based detection methods.

Table 5.28: **Equal Error Rate and operational performance comparison between *MADation* foundation model variants and *Gemma-3 12B-MAD*.** Evaluation includes FRLL morphing techniques, MIPGAN II, and MorDIFF datasets. BPCER values reported at fixed APCER thresholds of 1%, 10%, and 20%. Best results per metric highlighted in bold.

Dataset	Subset	MADation ViT-B				MADation ViT-L				Gemma3-12B-MAD (LoRA)			
		EER	BPCER@APCER			EER	BPCER@APCER			EER	BPCER@APCER		
		(%)	1%	10%	20%	(%)	1%	10%	20%	(%)	1%	10%	20%
FRLL	StyleGAN2	17.21	54.85	26.69	13.10	24.96	94.17	49.03	22.33	**6.83**	**22.06**	**5.39**	**3.43**
	WebMorph	3.42	5.88	**0.49**	**0.00**	4.07	6.86	1.47	1.47	**2.37**	**3.82**	**0.49**	0.49
	OpenCV	2.97	4.41	0.49	0.49	**0.99**	**0.98**	**0.00**	**0.00**	1.47	2.25	**0.00**	**0.00**
	AMSL	**3.85**	**12.07**	**2.89**	**2.41**	7.26	21.26	10.63	5.80	12.74	42.52	14.22	6.37
	FaceMorpher	1.35	1.47	**0.00**	**0.00**	0.74	0.98	0.98	0.98	**0.49**	**0.49**	**0.00**	**0.00**
MIPGAN II	–	22.21	84.80	47.55	26.47	**9.06**	100	**5.39**	**0.98**	17.92	55.00	23.57	15.29
MorDIFF	–	**1.10**	**1.94**	**0.00**	**0.00**	20.40	82.35	37.25	13.24	17.33	48.68	26.04	15.85
Average		**7.44**	**23.63**	11.16	6.07	9.64	43.80	14.96	6.40	8.45	24.97	**9.96**	**5.92**

Table 5.29: **Performance comparison on MIPGAN II dataset across multimodal large language models.** Results include *GPT-4 Turbo* (proprietary, zero-shot), *Gemma-3 27B* (open-source, zero-shot), and *Gemma-3 12B* variants with progressive adaptation strategies. Lower EER indicates better performance.

Model	MIPGAN II (EER %)
GPT4-Turbo [10]	37.0
Gemma3-27B Zero-Shot	35.56
Gemma3-12B Classification Head-Only	31.67
Gemma3-12B LoRA Fine-Tuned	**17.92**

Table 5.28 presents a direct comparison between *Gemma-3 12B-MAD* and *MADation*, the first vision-language foundation model adaptation for morphing attack detection. *MADation* employs CLIP's visual encoders (ViT-B and ViT-L) with lightweight LoRA fine-tuning, representing the most closely related approach to our methodology. Additionally, Table 5.29 compares our model variants against *GPT-4 Turbo*, providing context for multimodal LLM performance on challenging GAN-based morphs.

Foundation Model Architecture Comparison. *MADation ViT-B* achieves the lowest average EER at 7.44%, followed by *Gemma-3 12B-MAD* at 8.45% and *MADation ViT-L* at 9.64%. This ordering suggests that model scale does not directly correlate with detection performance; instead, the effectiveness depends on the interplay between architecture, pre-training, and adaptation strategy. *MADation ViT-B*'s superior average performance demonstrates that specialized vision-language models can be highly effective when properly adapted for morphing detection.

However, analysis at operational thresholds reveals a different picture. At 10% APCER, *Gemma-3 12B-MAD* achieves an average BPCER of 9.96%, outperforming both *MADation ViT-B* (11.16%) and ViT-L (14.96%). This superior operational performance extends to the 20% APCER threshold, where *Gemma-3* maintains the lowest average BPCER at 5.92%. These results indicate that while *MADation* may achieve better threshold-independent metrics, *Gemma-3* provides more reliable detection at security-critical operating points.

Performance by Morphing Technique. The models exhibit complementary strengths across different morphing types. For landmark-based morphs, all three approaches achieve exceptional performance, with sub-1% EER on FRLL-FaceMorpher and FRLL-OpenCV. *MADation ViT-L* achieves perfect detection (0.00% BPCER) on OpenCV morphs at 10% and 20% APCER, while *Gemma-3* demonstrates superior performance on FaceMorpher with 0.49% EER compared to ViT-B's 1.35% and ViT-L's 0.74%.

On StyleGAN2 morphs, *Gemma-3 12B-MAD* significantly outperforms both

MADation variants, achieving 6.83% EER versus 17.21% (ViT-B) and 24.96% (ViT-L). This advantage is particularly pronounced at operational thresholds, where *Gemma-3* maintains 5.39% BPCER at 10% APCER compared to 26.69% and 49.03% for *MADation* variants. This suggests that multimodal LLM architectures may possess superior capability for detecting sophisticated GAN artifacts.

The models show divergent performance on specialized datasets. *MADation ViT-B* excels on MorDIFF (1.10% EER), while ViT-L performs best on MIPGAN II (9.06% EER). *Gemma-3* achieves moderate performance on both (17.33% and 17.92% respectively), indicating that specific architectural choices within foundation models can optimize detection for particular morphing techniques.

Multimodal LLM Comparison. Table 5.29 provides crucial context by comparing *Gemma-3* variants against *GPT-4 Turbo* on the challenging MIPGAN II dataset. *GPT-4 Turbo*, despite being a significantly larger proprietary model, achieves only 37.0% EER in zero-shot evaluation. Our *Gemma-3 27B* zero-shot configuration slightly outperforms it at 35.56%, while the classification head-only baseline further improves to 31.67%. Most notably, LoRA fine-tuning reduces the EER to 17.92%, representing a 52% improvement over *GPT-4 Turbo*.

This dramatic performance gap demonstrates that model scale alone is insufficient for morphing detection; domain-specific adaptation is essential. The progression from zero-shot (35.56%) through classification head (31.67%) to LoRA fine-tuning (17.92%) illustrates how targeted adaptation progressively enhances detection capabilities, ultimately achieving performance competitive with specialized foundation models.

Adaptation Strategy Impact. Both *Gemma-3 12B-MAD* and *MADation* employ LoRA for parameter-efficient fine-tuning, but their base architectures and training strategies differ significantly. *MADation* adapts a pure vision model (CLIP) for classification, while *Gemma-3* leverages a multimodal LLM's combined vision-language understanding. The comparable performance between these approaches (7.44% vs. 8.45% average EER) validates that multiple foundation model architectures can be successfully adapted for morphing detection,

with each offering distinct advantages.

Key Findings. The evaluation establishes two critical insights.

First, foundation model adaptations (both *MADation* and *Gemma-3*) dramatically outperform larger models in zero-shot configuration, with our LoRA-tuned *Gemma-3* achieving **a 19-point EER improvement over *GPT-4 Turbo*** on *MIPGAN II*.

Second, while *MADation ViT-B* achieves the best average EER (7.44%), ***Gemma-3 12B-MAD* demonstrates superior operational performance** with lower BPCER at fixed APCER thresholds, suggesting better calibration for practical deployment.

5.3.4 Evaluation Summary

In this section, we summarize the comparative results from supervised, unsupervised, and foundation model baselines to provide an overall assessment of *Gemma-3 12B-MAD*'s performance position within the current state-of-the-art morphing attack detection landscape.

The comparative results position *Gemma-3 12B-MAD (LoRA)* as a consistently strong foundation-model approach that surpasses classical supervised detectors and competes closely with the best unsupervised/self-supervised and alternative foundation-model baselines. Against supervised deep networks trained under the *SPL-MAD* protocol, *Gemma-3* achieves the lowest average EER (14.81%), outperforming the best supervised average (*MixFaceNet-MAD* on SMDD, 21.55%). The gap widens at operational points: at 5% APCER, *Gemma-3*'s average BPCER is 18.13% versus 39.85% for the strongest supervised baseline, indicating materially better error trade-offs under security-relevant thresholds. Per-technique analysis shows especially large margins on landmark-based attacks (e.g., FRLL-OpenCV, FRLL-FaceMorpher, FRLL-WebMorph), while GAN-heavy sets such as MorGAN remain challenging for all supervised baselines and for *Gemma-3* alike.

When compared to unsupervised and self-supervised methods, *SelfMAD* remains the SOTA in average EER (5.74%), with *Gemma-3* ranking second (9.91%). However, the operational picture is more balanced. On several subsets, most notably FRLL-StyleGAN2, *Gemma-3* attains lower BPCER at fixed APCER than *SelfMAD* (e.g., at 10% APCER), indicating better calibration under deployment-like thresholds despite a higher threshold-independent error. Quality-heuristic baselines (*FIQA-MagFace*, *CNNIQA*) and earlier anomaly/self-paced models (*MAD-DDPM*, *SPL-MAD*) trail substantially on both averages and operating-point metrics, underscoring the advantage of either self-supervision with targeted artifacts (*SelfMAD*) on foundation-model adaptation (*Gemma-3*).

Among foundation-model adaptations, *MADation* (CLIP ViT-B/ViT-L) achieves the best average EER (7.44% for ViT-B), with *Gemma-3* close behind (8.45%) and ahead of *MADation ViT-L* (9.64%). Yet at fixed operating points, *Gemma-3* attains the lowest average BPCER at 10% and 20% APCER (9.96% and 5.92%, respectively), indicating superior operating-point efficiency. The methods exhibit complementary strengths: *Gemma-3* is dominant on FRLL-StyleGAN2 (6.83% EER vs. 17.21%/24.96% for *MADation* ViT-B/ViT-L), while *MADation* leads on AMSL and MorDIFF, and ViT-L is strongest on MIPGAN II. Finally, against a large multimodal LLM baseline, *GPT-4 Turbo* (zero-shot) records 37.0% EER on MIPGAN II; *Gemma-3* improves stepwise from 27B zero-shot (35.56%) to 12B head-only (31.67%), and to 12B LoRA (17.92%), demonstrating that task-specific adaptation, not model scale, is the key driver of performance.

Overall, the cross-baseline evidence shows: (i) clear wins over supervised MAD both on averages and at security-critical thresholds; (ii) near-SOTA standing vs. *SelfMAD*, with *Gemma-3* frequently superior at fixed APCER despite a higher average EER; and (iii) competitive parity with *MADation* on averages and advantages at operating points, particularly on StyleGAN-type attacks. Taken together, these results prove that a fine-tuned *Gemma-3 12B* is a highly capable morphing-attack detector, delivering competitive averages and superior operating-point behavior across several key benchmarks; this places *Gemma-3 12B-MAD* at the forefront of MAD, alongside *SelfMAD* and *MADation*.

6 Conclusion

This thesis presented a comprehensive investigation into the application of multimodal large language models for face morphing attack detection, establishing a new paradigm that leverages large language model capabilities for biometric security applications. Through systematic evaluation of open-source multimodal LLMs, development of sophisticated prompt engineering strategies, and implementation of parameter-efficient fine-tuning techniques, we demonstrated that general-purpose vision-language models can be successfully adapted to achieve competitive performance in morphing attack detection.

Our experimental results validate the core hypothesis that multimodal LLMs possess inherent forensic analysis capabilities that can be unlocked through appropriate instruction and adaptation. The zero-shot evaluation of four state-of-the-art models revealed that *Gemma-3 27B* achieved the best performance with an average EER of 32.09%, demonstrating measurable morphing detection ability without any task-specific training. This finding is significant as it establishes that foundation models trained on diverse internet-scale data implicitly learn visual patterns relevant to detecting facial manipulations.

The development of our structured forensic analysis framework represents a methodological contribution beyond simple prompt optimization. By guiding models through systematic six-step analytical procedures with fine-grained confidence scoring (0–10,000 scale), we achieved a 18% average EER reduction compared to basic prompting strategies. This framework not only improved detection accuracy but also provided interpretable, structured outputs that explain the reasoning behind each decision, a critical requirement for high-stakes security applications where transparency is essential.

The most substantial performance gains emerged from parameter-efficient fine-tuning using LoRA adaptation. Our fine-tuned *Gemma-3 12B* model achieved an average EER of 14.98%, representing a 53.3% improvement over zero-shot performance while requiring only 0.61% of trainable parameters. This approach successfully bridged the gap between general-purpose LLMs and specialized detection systems, achieving superior performance compared to supervised deep learning baselines (14.81% vs. 21.55% for the best baseline) and establishing the second-best results among all evaluated methods, surpassed only by the state-of-the-art *SelfMAD* approach (5.74% EER). Additionally, feature space analysis revealed that classification head calibration can become a bottleneck even when representational learning succeeds, as evidenced by *FERET-StyleGAN*'s excellent cluster separation yet poor EER performance. Furthermore, input resolution emerged as a critical factor transcending morphing methodology, as both GAN-based (*MorGAN-GAN*) and landmark-based (*MorGAN-LMA*) attacks at 64×64 resolution showed scattered feature distributions and failed genuine sample clustering, unlike the tight clustering patterns observed across all higher-resolution datasets.

Comparative analysis against existing methods revealed distinct advantages of the multimodal LLM approach. While *SelfMAD* remains superior in average performance, *Gemma-3 12B-MAD* demonstrated complementary strengths, particularly on specific GAN-based morphs where it outperformed *SelfMAD* on FRLL-StyleGAN2 (6.83% vs. 10.34% EER). Against other foundation model adaptations, our approach achieved comparable average performance to *MADation* (8.45% vs. 7.44% EER) while demonstrating superior operational characteristics at security-critical thresholds. Most notably, the 52% improvement over *GPT-4 Turbo* on MIPGAN II datasets conclusively demonstrates that targeted adaptation, rather than model scale alone, determines detection effectiveness.

From a practical deployment perspective, the fine-tuned *Gemma-3 12B* model offers compelling advantages. The 30× inference speedup compared to zero-shot evaluation (from 30 seconds to under 1 second per image) and the ability to deploy on a single GPU make it viable for real-world security systems.

6.1 Limitations

Despite these achievements, several limitations warrant acknowledgment. The performance gap with *SelfMAD* indicates that specialized self-supervised approaches remain superior for certain morphing types, particularly landmark-based attacks where *SelfMAD* achieves near-perfect detection. Our approach struggles with sophisticated GAN morphs like *MorGAN*, where EERs remain above 45%, suggesting that certain synthetic faces that closely approximate genuine distributions challenge even adapted foundation models. Additionally, the reliance on proprietary model architectures and substantial computational resources for training may limit accessibility for some research groups.

Feature analysis revealed classification head calibration issues where training data created decision boundaries poorly suited for certain real-world datasets, despite successful feature separation. Additionally, input resolution critically impacts detection capability regardless of morphing technique, as low-resolution datasets (64×64 pixels) prevented the model from establishing reliable feature representations for both genuine samples and attacks, leading to scattered embeddings resembling untrained states.

6.2 Future Directions

This research opens the door for future LLM adaptation and implementation for the MAD task. Different fine-tuning approaches could be explored, particularly those that explicitly teach the model to provide detailed reasoning behind its decisions, potentially resolving the critical issue of explainability in morphing attack detection. Rather than training solely on classification objectives, future work could leverage the generative capabilities of LLMs to produce comprehensive forensic reports that detail specific artifacts and their locations, enhancing both detection accuracy and interpretability.

Investigating improved domain adaptation strategies could address classification calibration issues, particularly developing training protocols that ensure decision boundaries generalize across diverse imaging conditions and dataset char-

acteristics. Additionally, establishing minimum resolution requirements for effective morphing detection could guide preprocessing standards and dataset creation protocols for future *LLM-MAD* research.

Additionally, the adaptation of smaller multimodal LLMs could reveal the minimum model capacity required for effective morphing detection, potentially enabling deployment on resource-constrained devices. Conversely, exploring larger LLMs with enhanced reasoning capabilities could determine whether superior reasoning abilities translate to improved detection performance, particularly for sophisticated attacks that current models struggle to identify. The relationship between model scale, reasoning capability, and morphing detection accuracy remains an open question with significant implications for practical deployment strategies.

6.3 Final Remarks

This thesis demonstrates that multimodal large language models represent a viable and promising approach for morphing attack detection, achieving competitive performance with established methods while offering unique advantages in interpretability, deployment efficiency, and generalization capability. By establishing that general-purpose large language models can be successfully adapted for specialized security tasks, this work contributes to the broader understanding of how large-scale pre-training can benefit domain-specific applications. As morphing techniques continue to evolve and new attack vectors emerge, the flexibility and adaptability of foundation model approaches position them as valuable tools in the ongoing effort to secure biometric systems against sophisticated presentation attacks.

Bibliography

[1] M. Ivanovska, L. Todorov, N. Damer, D. K. Jain, P. Peer, and V. Štruc, "Selfmad: Enhancing generalization and robustness in morphing attack detection via self-supervised learning," *arXiv preprint arXiv:2504.05504*, 2025.

[2] M. Ferrara, A. Franco, and D. Maltoni, *On the Effects of Image Alterations on Face Recognition Accuracy.* Cham: Springer International Publishing, 2016, pp. 195–222. [Online]. Available: https://doi.org/10.1007/978-3-319-28501-6_9

[3] M. Huber, F. Boutros, A. T. Luu, K. B. Raja, R. Ramachandra, N. Damer, P. C. Neto, T. Gonçalves, A. F. Sequeira, J. S. Cardoso *et al.*, "Syn-mad 2022: Competition on face morphing attack detection based on privacy-aware synthetic training data," in *2022 International Joint Conference on Biometrics (IJCB).* IEEE, 2022, pp. 1–10.

[4] E. Caldeira, G. Ozgur, T. Chettaoui, M. Ivanovska, P. Peer, F. Boutros, V. Struc, and N. Damer, "Madation: Face morphing attack detection with foundation models," in *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 1650–1660.

[5] N. Damer, N. Spiller, M. Fang, F. Boutros, F. Kirchbuchner, and A. Kuijper, "Pw-mad: Pixel-wise supervision for generalized face morphing attack detection," in *International Symposium on Visual Computing (ISVC).* Springer, 2021, pp. 291–304.

[6] M. Fang, F. Boutros, and N. Damer, "Unsupervised face morphing attack detection via self-paced anomaly detection," in *2022 IEEE International Joint Conference on Biometrics (IJCB).* IEEE, 2022, pp. 1–11.

[7] R. Shekhawat, H. Li, R. Ramachandra, and S. Venkatesh, "Towards zero-shot differential morphing attack detection with multimodal large language models," *arXiv preprint arXiv:2505.15332*, 2025.

[8] E. Caldeira, F. Boutros, and N. Damer, "Madprompts: Unlocking zero-shot morphing attack detection with multiple prompt aggregation," *arXiv preprint arXiv:2508.08939*, 2025.

[9] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, "Face recognition systems under morphing attacks: A survey," *IEEE Access*, vol. 7, pp. 23 012–23 026, 2019.

[10] H. Zhang, R. Ramachandra, K. Raja, and C. Busch, "Chatgpt encounters morphing attack detection: Zero-shot mad with multi-modal large language models and general vision models," *arXiv preprint arXiv:2503.10937*, 2025.

[11] P. C. Neto, T. Gonçalves, M. Huber, N. Damer, A. F. Sequeira, and J. S. Cardoso, "Orthomad: Morphing attack detection through orthogonal identity disentanglement," in *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2022, pp. 1–5.

[12] M. Hamza, S. Tehsin, M. Humayun, M. F. Almufareh, and M. Alfayad, "A comprehensive review of face morph generation and detection of fraudulent identities," *Applied Sciences*, vol. 12, no. 24, p. 12545, 2022.

[13] E. Caldeira, P. C. Neto, T. Gonçalves, N. Damer, A. F. Sequeira, and J. S. Cardoso, "Unveiling the two-faced truth: Disentangling morphed identities for face morphing detection," in *2023 31th European Signal Processing Conference (EUSIPCO)*. IEEE, 2023.

[14] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[15] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff *et al.*, "The prompt report: a systematic survey of prompt engineering techniques," *arXiv preprint arXiv:2406.06608*, 2024.

[16] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *International Conference on Learning Representations (ICLR)*, vol. 1, no. 2, p. 3, 2022.

[17] Gemma Team, Google DeepMind, "Gemma-3-27B-IT," https://huggingface.co/google/gemma-3-27b-it, 2025, instruction-tuned multimodal model (27B parameters), released March 2025; accessed July 2025.

[18] Qwen Team, Alibaba Cloud, "Qwen2.5-vl-32b-instruct," https://huggingface.co/Qwen/Qwen2.5VL32BInstruct, 2025, instruction-tuned vision-language model, released March 24 2025 under Apache 2.0; accessed July 2025.

[19] Meta AI (Meta Platforms), "Llama-4-Scout-17B-16E-Instruct," https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct, 2025, instruction-tuned multimodal model (17B active parameters across 16 experts, 109B total), released April 5 2025; accessed August 4 2025.

[20] Mistral AI, "Mistral-small-3.1-24b-instruct-2503," https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503, 2025, instruction-tuned multimodal model with 24B parameters, 128k context window, and vision capabilities; released March 13, 2025 under Apache 2.0 license; accessed August 4, 2025.

[21] F. Boutros, N. Damer, M. Fang, F. Kirchbuchner, and A. Kuijper, "Mixfacenets: Extremely efficient face recognition networks," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2021, pp. 1–8.

[22] L. DeBruine and B. Jones, "Face research lab london set," May 2017. [Online]. Available: https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666/3

[23] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 947–954.

[24] P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss, "The feret database and evaluation procedure for face-recognition algorithms," *Image and vision computing*, vol. 16, no. 5, pp. 295–306, 1998.

[25] Z. W. Blasingame and C. Liu, "Greedy-dim: Greedy algorithms for unreasonably effective face morphs," in *2024 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, Sep. 2024, p. 1–11. [Online]. Available: http://dx.doi.org/10.1109/IJCB62174.2024.10744517

[26] H. Zhang, S. Venkatesh, R. Ramachandra, K. B. Raja, N. Damer, and C. Busch, "Mipgan: Generating strong and high quality morphing attacks using identity prior driven gan," *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*, vol. 3, no. 3, pp. 365–383, 2021.

[27] N. Damer, A. Moseguí Saladié, A. Braun, and A. Kuijper, "Morgan: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–10. [Online]. Available: https://doi.org/10.1109/BTAS.2018.8698563

[28] N. Damer, C. A. F. López, M. Fang, N. Spiller, M. V. Pham, and F. Boutros, "Privacy-friendly synthetic data for the development of face morphing attack detectors," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2022, New Orleans, LA, USA, June 19-20, 2022*. IEEE, 2022, pp. 1605–1616. [Online]. Available: https://doi.org/10.1109/CVPRW56347.2022.00167

[29] N. Damer, M. Fang, P. Siebke, J. N. Kolf, M. Huber, and F. Boutros, "Mordiff: Recognition vulnerability and attack detectability of face morphing attacks created by diffusion autoencoders," *arXiv preprint arXiv:2302.01843*, 2023. [Online]. Available: https://arxiv.org/abs/2302.01843

[30] H. Zhang, R. Ramachandra, K. Raja, and C. Busch, "Morph-pipe: Plugging in identity prior to enhance face morphing attack based on diffusion model," in *Proceedings of the Norwegian Information Security Conference (NISK)*, 2023.

[31] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[32] J. Kannala and E. Rahtu, "Bsif: Binarized statistical image features," in *Proceedings of the 21st international conference on pattern recognition (ICPR2012)*. IEEE, 2012, pp. 1363–1366.

[33] R. Ramachandra, K. Raja, and C. Busch, "Detecting morphed face images," in *8th IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2016, pp. 1–7.

[34] L. Debiasi, C. Rathgeb, U. Scherhag, A. Uhl, and C. Busch, "Prnu variance analysis for morphed face image detection," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–9.

[35] L.-B. Zhang, F. Peng, and M. Long, "Face morphing detection using fourier spectrum of sensor pattern noise," in *2018 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2018, pp. 1–6.

[36] U. Scherhag, L. Debiasi, C. Rathgeb, C. Busch, and A. Uhl, "Detection of face morphing attacks based on prnu analysis," *IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM)*, vol. 1, no. 4, pp. 302–317, 2019.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[39] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 2818–2826.

[40] M. Ivanovska and V. Štruc, "Face morphing attack detection with denoising diffusion probabilistic models," in *2023 11th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2023, pp. 1–6.

[41] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2021, pp. 14 225–14 234.

[42] P. Terhörst, J. N. Kolf, N. Damer, F. Kirchbuchner, and A. Kuijper, "Face quality estimation and its correlation to demographic and non-demographic bias in face recognition," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020, pp. 1–11.

[43] J. Hernandez-Ortega, J. Galbally, J. Fierrez, R. Haraksim, and L. Beslay, "Faceqnet: Quality assessment for face recognition based on deep learning," in *2019 International Conference on Biometrics (ICB)*. IEEE, 2019, pp. 1–8.

[44] J. Chen, Y. Deng, G. Bai, and G. Su, "Face image quality assessment based on learning to rank," *IEEE signal processing letters*, vol. 22, no. 1, pp. 90–94, 2014.

[45] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2014, pp. 1733–1740.

[46] W. Shao and X. Mou, "No-reference image quality assessment based on edge pattern feature in the spatial domain," *IEEE Access*, vol. 9, pp. 133 170–133 184, 2021.

[47] N. Venkatanath, D. Praneeth, S. C. Sumohana, S. M. Swarup *et al.*, "Blind image quality evaluation using perception based features," in *2015 twenty first national conference on communications (NCC)*. IEEE, 2015, pp. 1–6.

[48] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal processing letters*, vol. 20, no. 3, pp. 209–212, 2012.

[49] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Transactions on image processing (TIP)*, vol. 27, no. 1, pp. 206–219, 2017.

[50] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Transactions on image processing (TIP)*, vol. 26, no. 8, pp. 3951–3964, 2017.

[51] B. Fu and N. Damer, "Face morphing attacks and face image quality: The effect of morphing and the unsupervised attack detection by quality," *IET Biometrics*, vol. 11, no. 5, pp. 359–382, 2022. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/bme2.12094

[52] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8110–8119.

[53] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[54] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière *et al.*, "Gemma 3 technical report," *arXiv preprint arXiv:2503.19786*, 2025.

[55] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.

[56] Mistral AI, "Mistral small 3.1 (mistral-small-2503) instruct," https://mistral.ai/news/mistral-small-3-1, 2025, released March 17 2025 under Apache 2.0, 24B parameters, 128k context window, multimodal (text vision); accessed August 4 2025.

[57] H. Zhao, Z. Cai, S. Si, X. Ma, K. An, L. Chen, Z. Liu, S. Wang, W. Han, and B. Chang, "Mmicl: Empowering vision-language model with multi-modal in-context learning," in *International Conference on Learning Representations (ICLR)*, 2024.

[58] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[59] Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang, "An empirical study of catastrophic forgetting in large language models during continual fine-tuning," *arXiv preprint arXiv:2308.08747*, 2023.

[60] *Information technology — Biometric presentation attack detection — Part 3: Testing and reporting*, International Organization for Standardization Std. ISO/IEC 30 107-3:2023, 2023. [Online]. Available: https://www.iso.org/standard/79520.html

[61] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel, "Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks," *arXiv preprint*, Oct. 2020. [Online]. Available: https://arxiv.org/abs/2012.05344

[62] ——, "Are gan-based morphs threatening face recognition?" in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 2959–2963. [Online]. Available: https://doi.org/10.1109/ICASSP43922.2022.9746477

[63] L. DeBruine, "webmorph," 2018. [Online]. Available: https://github.com/debruine/webmorph

[64] T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, and J. Dittmann, "Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images," *IET Biometrics*, vol. 7, no. 4, pp. 325–332, 2018. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-bmt.2017.0147

[65] M. Ivanovska, A. Kronovšek, P. Peer, V. Štruc, and B. Batagelj, "Face morphing attack detection using privacy-aware training data," *arXiv preprint arXiv:2207.00899*, 2022.

[66] K. Shoemake, "Animating rotation with quaternion curves," in *Proceedings of the 12th annual conference on Computer graphics and interactive techniques (SIGGRAPH)*. ACM, 1985, pp. 245–254.

[67] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=St1giarCHLP

[68] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 6840–6851.

[69] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 609–10 619.

[70] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "Elasticface: Elastic margin loss for deep face recognition," *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1577–1586, 2022.

[71] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4690–4699.

[72] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 155–11 164. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Kim_AdaFace_Quality_Adaptive_Margin_for_Face_Recognition_CVPR_2022_paper.html

[73] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

[74] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 3730–3738.

[75] International Civil Aviation Organization (ICAO), "Portrait quality: Reference facial images for mrtd," ICAO, Tech. Rep., 2017, technical Report, Version 0.9.

[76] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 67–74.

[77] International Civil Aviation Organization (ICAO), "Machine readable travel documents, part 9: Deployment of biometric identification and electronic storage of data in emrtds," ICAO, Tech. Rep., 2015, iCAO Doc 9303, Part 9.

[78] R. Raghavendra, K. B. Raja, S. Venkatesh, and C. Busch, "Face morphing versus face averaging: Vulnerability and detection," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 555–563.

[79] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," Carnegie Mellon University, School of Computer Science, Tech. Rep. CMU-CS-16-118, 2016. [Online]. Available: https://cmusatyalab.github.io/openface/

[80] Frontex, "Best practice technical guidelines for automated border control (abc) systems," European Agency for the Management of Operational Cooperation at the External Borders of the Member States of the European Union, Tech. Rep., 2015.

[81] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[82] F. Boutros, M. Fang, M. Klemt, B. Fu, and N. Damer, "Cr-fiqa: face image quality assessment by learning sample relative classifiability," in *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 5836–5845.

[83] S. Mallick, "Face morph using opencv — c++ / python," http://www.learnopencv.com/face-morph-using-opencv-cpp-python/, 2016.

[84] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1867–1874.

[85] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.

[86] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica, "Efficient memory management for large language model serving with pagedattention," in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[87] Z. Li, Y. Su, R. Yang, C. Xie, Z. Wang, Z. Xie, N. Wong, and H. Yang, "Quantization meets reasoning: Exploring llm low-bit quantization degradation for mathematical reasoning," *arXiv preprint arXiv:2501.03035*, 2025.

[88] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 720–18 729.

Appendix

A Complete Prompt Texts

A.1 Prompt 1: Structured Forensic Analysis – Semantic Guide A

Listing A.1: Prompt 1: Structured Forensic Analysis – Semantic Guide A

```
["""# Multimodal Morphing Attack Detection Prompt

## Task
You are a multimodal AI system with both **vision and language understanding**. You are given an image of a face and your task is to determine whether it is a **bona fide (authentic)** face or a **morphing attack** (a digitally blended face of two people).

After each step of your analysis, you must provide a **confidence score** on a strict scale from **0.000 to 10000.000**, where:

- `0.000` means **definitely bona fide**
- `10000.000` means **definitely morph**
- Values in between reflect **fine-grained probability** of a morphing attack
- 0-1000: Strong evidence of authentic face
- 1000-3000: Likely authentic with minor irregularities
- 3000-7000: Uncertain, requires careful analysis
- 7000-9000: Likely morphed with moderate evidence
- 9000-10000: Strong evidence of morphing attack
```

```
16
17 > **Important:** Avoid rounded or generic values like
       `1000`, `5000`, `8000`. Your score must be **precise**,
       with at least **three decimal places**. This is
       essential for biometric performance analysis and
       threshold calibration.
18
19 Use the **step-by-step visual analysis** outlined below.
       After each step, provide a confidence score in the
       specified format, based on your observations and the
       responses to the guiding questions.
20
21 ## Visual Analysis Steps
22
23 ### Step 1: Core Facial Features
24 - Focus on the **eyes, nose, lips, and eyebrows**.
25 - Look for signs of **ghosting**, **faint duplicates**, or
       misaligned or unnatural elements.
26 - Check if eye contours or lip lines appear duplicated or
       semi-transparent.
27
28 **Ask**: "Do facial features have any doubled contours or
       blended boundaries?"
29 **Ask**: "Do the eyes appear blurred or duplicated?"
30 **Ask**: "Do the lips show any visual artifacts, or are the
        lip lines irregular?"
31
32 After this step, provide a confidence score for Step 1 in
       this format:
33 ```json
34 {"step1_score": [0 to 10000]}
35 ```
36
37 ### Step 2: Facial Geometry and Symmetry
38 - Visually compare the **left and right halves** of the
       face.
```

```
- Detect any asymmetry in shape, spacing, or size of eyes, irises, ears, and jawline.
- Assess if the overall geometry seems subtly misaligned or "averaged."

**Ask**: "Do the facial proportions look unnaturally blended or off-balance?"

After this step, provide a confidence score for Step 2 in this format:
```json
{"step2_score": [0 to 10000]}
```

Step 3: Skin Texture and Detail
- Inspect the **skin surface** for fine detail.
- Detect over-smoothness, uniform skin tone, or "plastic-like" appearance.
- Check if pores, wrinkles, or blemishes are abnormally absent or symmetric.

Ask: "Does the skin look too perfect, synthetic, or even-textured?"
Ask: "Are pores, wrinkles, or blemishes abnormally absent or symmetric?"

After this step, provide a confidence score for Step 3 in this format:
```json
{"step3_score": [0 to 10000]}
```

Step 4: Image Boundary and Hairline
- Look at the **face boundary**, including **ears, hairline, and background**.
- Detect any blurred transitions, blending seams, or edge mismatches.
```

```
- Check for faded or semi-transparent features outside the main face.

**Ask**: "Do any facial borders blend unnaturally into the background?"
**Ask**: "Are there any faded or semi-transparent elements present in the image?"
**Ask**: "Do ears, hairline, and background have any faded or semi-transparent artifacts?"

After this step, provide a confidence score for Step 4 in this format:
```json
{"step4_score": [0 to 10000]}
```

Step 5: Lighting and Color Consistency
- Examine **lighting direction, reflections, and shadows**.
- Identify **saturation anomalies** (e.g., bright red patches) or abnormal gradients.
- Compare lighting consistency across both sides of the face.

Ask: "Are there unnatural color shifts or inconsistent lighting effects?"

After this step, provide a confidence score for Step 5 in this format:
```json
{"step5_score": [0 to 10000]}
```

Step 6: Identity Coherence
- Assess whether all facial features belong to the **same plausible identity**.
- Detect any mismatch (e.g., eyes that don't match the rest of the face).
```

```
- Evaluate whether the face looks **overly generic or statistically averaged**.

**Ask**: "Does the face appear to be a convincing real person, or a blend of two?"

After this step, provide a confidence score for Step 6 in this format:
```json
{"step6_score": [0 to 10000]}
```

Final Analysis Summary

After completing all steps, summarize your overall findings and indicate which steps showed the strongest evidence of a potential morphing attack, if any.
"""
]
```

A.2 Prompt 2: Extended Forensic Analysis – Semantic Guide A

Listing A.2: Prompt 2: Extended Forensic Analysis – Semantic Guide A

```
["""# Multimodal Morphing Attack Detection Prompt

## Task
You are a multimodal AI system with both **vision and language understanding**. You are given an image of a face and your task is to determine whether it is a **bona fide (authentic)** face or a **morphing attack** (a digitally blended face of two people).
Use your visual understanding capabilities to examine the image carefully.
```

```
6  Take time to analyze each aspect before providing your
       confidence score.
7
8  After each step of your analysis, you must provide a **
       confidence score** on a strict scale from **0.000 to
       10000.000**, where:
9
10 - `0.000` means **definitely bona fide**
11 - `10000.000` means **definitely morph**
12 - Values in between reflect **fine-grained probability** of
        a morphing attack
13 - 0-1000: Strong evidence of authentic face
14 - 1000-3000: Likely authentic with minor irregularities
15 - 3000-7000: Uncertain, requires careful analysis
16 - 7000-9000: Likely morphed with moderate evidence
17 - 9000-10000: Strong evidence of morphing attack
18
19
20 > **Scoring Guidelines:**
21 > - Consider the severity and frequency of artifacts
22 > - Multiple minor issues may accumulate to higher scores
23 > - A single clear morphing artifact should result in a
       high score (7000+)
24 > - When in doubt, err on the side of caution but justify
       your reasoning
25
26 > **Important:** Avoid rounded or generic values like
       `1000`, `5000`, `8000`. Your score must be **precise**,
       with at least **three decimal places**. This is
       essential for biometric performance analysis and
       threshold calibration.
27
28 Use the **step-by-step visual analysis** outlined below.
       After each step, provide a confidence score in the
       specified format, based on your observations and the
       responses to the guiding questions.
29
```

```
## Confidence Calibration
Before providing each score, consider:
- How certain are you about this assessment?
- Have you seen similar patterns in legitimate faces?
- Would a human expert notice these artifacts?

## Visual Analysis Steps

### Step 1: Core Facial Features
- Focus on the **eyes, nose, lips, and eyebrows**.
- Look for signs of **ghosting**, **faint duplicates**, or misaligned or unnatural elements.
- Check if eye contours or lip lines appear duplicated or semi-transparent.

**Ask**: "Do facial features have any doubled contours or blended boundaries?"
**Ask**: "Do the eyes appear blurred or duplicated?"
**Ask**: "Do the lips show any visual artifacts, or are the lip lines irregular?"

After this step, provide a confidence score for Step 1 in this format:
```json
{"step1_score": [0 to 10000]}
```

Step 2: Facial Geometry and Symmetry
- Visually compare the **left and right halves** of the face.
- Detect any asymmetry in shape, spacing, or size of eyes, irises, ears, and jawline.
- Assess if the overall geometry seems subtly misaligned or "averaged."

Ask: "Do the facial proportions look unnaturally blended or off-balance?"
```

```
**Ask**: "Is there any asymmetry in shape, spacing, or size of eyes, irises, ears, and jawline?"
**Ask**: "Does the overall facial geometry appear artificially averaged or unnaturally symmetric?"

After this step, provide a confidence score for Step 2 in this format:
```json
{"step2_score": [0 to 10000]}
```

Step 3: Skin Texture and Detail
- Inspect the **skin surface** for fine detail.
- Detect over-smoothness, uniform skin tone, or "plastic-like" appearance.
- Check if pores, wrinkles, or blemishes are abnormally absent or symmetric.

Ask: "Does the skin look too perfect, synthetic, or even-textured?"
Ask: "Are pores, wrinkles, or blemishes abnormally absent or symmetric?"

After this step, provide a confidence score for Step 3 in this format:
```json
{"step3_score": [0 to 10000]}
```

Step 4: Image Boundary and Hairline
- Look at the **face boundary**, including **ears, hairline, and background**.
- Detect any blurred transitions, blending seams, or edge mismatches.
- Check for faded or semi-transparent features outside the main face.

```

```
Ask: "Do any facial borders blend unnaturally into the background?"
Ask: "Are there any faded or semi-transparent elements present in the image?"
Ask: "Do ears, hairline, and background have any faded or semi-transparent artifacts?"

After this step, provide a confidence score for Step 4 in this format:
```json
{"step4_score": [0 to 10000]}
```

Step 5: Lighting and Color Consistency
- Examine **lighting direction, reflections, and shadows**.
- Identify **saturation anomalies** (e.g., bright red patches) or abnormal gradients.
- Compare lighting consistency across both sides of the face.

Ask: "Are there unnatural color shifts or inconsistent lighting effects?"
Ask: "Are there any saturation anomalies?"
Ask: "Is lighting consistent across both sides of the face?"

After this step, provide a confidence score for Step 5 in this format:
```json
{"step5_score": [0 to 10000]}
```

Step 6: Identity Coherence
- Assess whether all facial features belong to the **same plausible identity**.
- Look for inconsistencies that suggest multiple source faces
```

```
- Evaluate whether the face appears artificially averaged or generic
- Check if features match in terms of age, ethnicity, and style

**Ask**: "Do all facial features appear to belong to the same person?"
**Ask**: "Does the face look artificially perfect or overly averaged?"
**Ask**: "Are there any inconsistencies in age, ethnicity, or style between features?"

After this step, provide a confidence score for Step 6 in this format:
```json
{"step6_score": [0 to 10000]}
```

Final Analysis Summary

After completing all steps, summarize your overall findings and indicate which steps showed the strongest evidence of a potential morphing attack, if any.
"""]
```

A.3 Prompt 3: Optimized Forensic Analysis – Semantic Guide B

Listing A.3: Prompt 3: Optimized Forensic Analysis – Semantic Guide B

```
["""
**Objective:**
Act as a forensic image analysis expert. Your primary task is to **quantify the level of suspicion** that a given facial image is a **morphing attack**. You will analyze the image for subtle artifacts and indicators, assigning
```

```
    a precise score that reflects your level of confidence.

> **Guiding Principle: From Deterministic to Probabilistic Analysis**
> Your goal is **not** to simply classify the image as "bona fide" or "morph." Instead, you must analyze the visual evidence and quantify your suspicion on a fine-grained spectrum. Acknowledge ambiguity. A lack of obvious artifacts does not automatically mean a score of 0, nor does a single minor anomaly warrant a score of 10000. Your analysis must produce scores that utilize the **full range** of the scale, reflecting the subtle nature of morphing attacks.

**Scoring Mandate & Semantic Guide:**
You MUST use the entire **0.000 to 10000.000** scale. Scores must have **three decimal places**. Use the following guide to map your findings to a score:

-   **‘0.000 - 1000.000‘ (Very Low Suspicion):** Image appears clean, coherent, and authentic. No significant artifacts detected. Corresponds to high confidence in authenticity.
-   **‘1000.001 - 4000.000‘ (Low to Moderate Suspicion):** One or two minor, inconclusive artifacts are present (e.g., slight unnatural smoothness, minor asymmetry). These could potentially be explained by compression, lighting, or natural features, but warrant a degree of suspicion.
-   **‘4000.001 - 6000.000‘ (Ambiguous / Moderate Suspicion):** There are noticeable artifacts that are suspicious, but no single piece of evidence is conclusive. The image feels "off." This is the zone of highest uncertainty.
-   **‘6000.001 - 9000.000‘ (High Suspicion):** Multiple, distinct artifacts are present across different areas of the face (e.g., ghosting on eyes, edge blurring, and
```

```
   inconsistent lighting). A morph is highly probable.
15 -   **‘9000.001 - 10000.000‘ (Very High / Near Certainty)
   :** Overwhelming and clear evidence of morphing.
   Multiple, strong artifacts are easily identifiable and
   create an incoherent image.
16
17 ---
18
19 ### Forensic Analysis Steps
20
21 **Step 1: Core Facial Feature Analysis**
22 -   **Area of Focus:** Eyes, nose, lips, and eyebrows.
23 -   **Artifacts to Detect:** Ghosting, faint duplicates,
   misaligned elements, doubled contours, unnaturally
   blurred eye contours, or irregular/asymmetric lip lines.
24
25 **Step 2: Facial Geometry and Symmetry Analysis**
26 -   **Area of Focus:** Overall facial structure, comparing
   the left and right halves.
27 -   **Artifacts to Detect:** Unnatural asymmetry in the
   size or spacing of eyes, irises, or ears. A jawline that
    appears averaged or ill-defined. Proportions that seem
   subtly misaligned or blended.
28
29 **Step 3: Skin Texture and Detail Analysis**
30 -   **Area of Focus:** Skin surface across the entire face
   (cheeks, forehead, chin).
31 -   **Artifacts to Detect:** Overly smooth, "plastic-like,"
    or synthetic skin texture. An abnormal lack of fine
   details like pores, micro-wrinkles, or minor blemishes.
   Unnatural patterns or symmetry in skin details.
32
33 **Step 4: Boundary and Edge Analysis**
34 -   **Area of Focus:** The outer perimeter of the face,
   including the hairline, ears, and jawline transitioning
   into the background.
```

```
- **Artifacts to Detect:** Blurring, smudging, or seam-like artifacts at the edge of the face. Faded or semi-transparent features, especially around the ears or hair. Inconsistencies between the focus/sharpness of the face and the background.

**Step 5: Lighting and Color Consistency Analysis**
- **Area of Focus:** The entire image, paying attention to light, shadow, and color.
- **Artifacts to Detect:** Inconsistent lighting direction, unnatural color shifts, mismatched lighting on different parts of the face, or inconsistent specular highlights in the eyes.

**Step 6: Identity Coherence Analysis**
- **Area of Focus:** The holistic impression of the face as a single, plausible identity.
- **Artifacts to Detect:** Features that appear to belong to different people, an "overly generic" or statistically averaged appearance, or a general feeling that the face is not a convincing, real person.

---

### Final Output Requirement

After completing your six-step analysis, present your complete findings in a single JSON object. Do not provide any text or explanation outside of this JSON block.

**JSON Format:**
‘‘‘json
{
  "final_decision": {
    "overall_confidence_score": [Value between 0.000 and 10000.000],
```

```
    "summary_of_findings": "A brief summary justifying the
       overall score, referencing the Semantic Scoring
       Guide and highlighting the key evidence (or lack
       thereof)."
  },
  "step_by_step_analysis": {
    "step1_core_features": {
      "score": [Value between 0.000 and 10000.000],
      "rationale": "Describe observed artifacts and explain
          why the score reflects a specific level of
        suspicion (e.g., 'Faint asymmetry noted in lip
        corners, leading to a low-suspicion score of
        1850.455')."
    },
    "step2_facial_geometry": {
      "score": [Value between 0.000 and 10000.000],
      "rationale": "Describe observed artifacts and explain
          why the score reflects a specific level of
        suspicion."
    },
    "step3_skin_texture": {
      "score": [Value between 0.000 and 10000.000],
      "rationale": "Describe observed artifacts and explain
          why the score reflects a specific level of
        suspicion."
    },
    "step4_boundaries_and_edges": {
      "score": [Value between 0.000 and 10000.000],
      "rationale": "Describe observed artifacts and explain
          why the score reflects a specific level of
        suspicion."
    },
    "step5_lighting_and_color": {
      "score": [Value between 0.000 and 10000.000],
      "rationale": "Describe observed artifacts and explain
          why the score reflects a specific level of
        suspicion."
```

```
    },
    "step6_identity_coherence": {
      "score": [Value between 0.000 and 10000.000],
      "rationale": "Describe observed artifacts and explain
          why the score reflects a specific level of
         suspicion."
    }
  }
}
```
"""]