



Guard Me If You Know Me: Protecting Specific Face-Identity from Deepfakes

Kaiqing Lin^{1,3*}, Zhiyuan Yan^{2,3*}, Ke-Yue Zhang³, Li Hao², Yue Zhou¹, Yuzhen Lin¹,
Weixiang Li¹, Taiping Yao^{3†}, Shouhong Ding³, Bin Li^{1†}

¹Guangdong Provincial Key Laboratory of Intelligent Information Processing,
Shenzhen Key Laboratory of Media Security,
and SZU-AFS Joint Innovation Center for AI Technology, Shenzhen University

²School of Electronic and Computer Engineering, Peking University,

³Tencent Youtu Lab

Abstract

Securing personal identity against deepfake attacks is increasingly critical in the digital age, especially for celebrities and political figures whose faces are easily accessible and frequently targeted. Most existing deepfake detection methods focus on general-purpose scenarios and often ignore the valuable prior knowledge of known facial identities, e.g., "VIP individuals" whose authentic facial data are already available. In this paper, we propose **VIPGuard**, a unified multimodal framework designed to capture fine-grained and comprehensive facial representations of a given identity, compare them against potentially fake or similar-looking faces, and reason over these comparisons to make accurate and explainable predictions. Specifically, our framework consists of three main stages. First, we fine-tune a multimodal large language model (MLLM) to learn detailed and structural facial attributes. Second, we perform identity-level discriminative learning to enable the model to distinguish subtle differences between highly similar faces, including real and fake variations. Finally, we introduce user-specific customization, where we model the unique characteristics of the target face identity and perform semantic reasoning via MLLM to enable personalized and explainable deepfake detection. Our framework shows clear advantages over previous detection works, where traditional detectors mainly rely on low-level visual cues and provide no human-understandable explanations, while other MLLM-based models often lack a detailed understanding of specific face identities. To facilitate the evaluation of our method, we build a comprehensive identity-aware benchmark called **VIPBench** for personalized deepfake detection, involving the latest 7 face-swapping and 7 entire face synthesis techniques for generation. Extensive experiments show that our model outperforms existing methods in both detection and explanation. The code is available at <https://github.com/KQL11/VIPGuard>.

1 Introduction

The rapid advancement of generative AI techniques [51, 59, 68, 75, 52, 13] has led to the widespread creation and dissemination of deepfake content—synthetic media where a person's identity is manipulated or replaced, often without consent. These techniques severely threaten personal reputation and

* Equal contribution, † Corresponding author

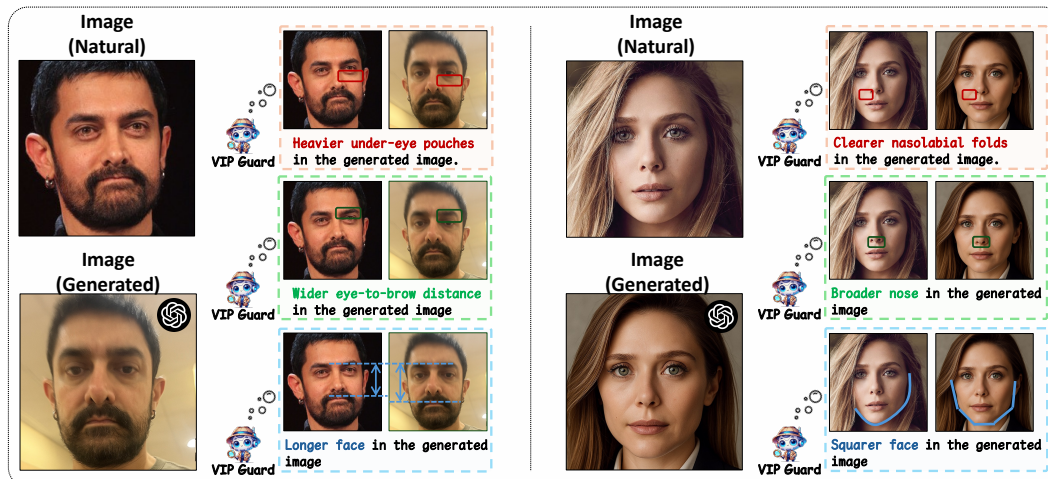


Figure 1: An illustrative comparison between natural images (from LAION-Face [87]) and images generated by GPT-4o [48], showing localized inconsistencies in facial attributes, such as eye pouches and facial shapes.

public trust, especially for high-profile individuals such as celebrities, political figures, and public officials [2]. To this end, protecting one's identity from such manipulations has become more than just a personal privacy issue—it is a pressing societal concern.

Although deepfake detection has gained attention, most existing methods are designed for general-purpose use [50, 36, 56, 76, 72, 44, 88, 89, 82, 83, 84]. They aim to classify *any* face image or video as real or fake, without considering *whose* face is being targeted. In the real world, many scenarios provide access to prior knowledge about the target identity¹, such as in the case of public figures or known individuals [16]. This opens a new avenue for detection: *Can we leverage the known facial identity to improve both detection and explainability in personalized deepfake detection?* Rather than treating all faces the same, identity-aware detection methods [54, 19, 17, 16, 81] focus on the semantic alignment between the input image and the authentic identity. By doing so, they provide more personalized and context-aware predictions, potentially improving both detection performance and explainability.

However, existing identity-aware detectors [16, 81] fail to fully utilize the detailed identity-specific information. They primarily rely on global facial features while neglecting fine-grained semantic details—such as eye shape, facial contours, or other attributes. As shown in Figure 1, an image generated by the latest GPT-4o [48] can appear highly realistic at first glance, but still contains subtle inconsistencies in local facial regions—for example, unusually pronounced eye bags. When the target identity is known and all facial details are available, leveraging these fine-grained discrepancies for detection becomes especially promising.

To this end, in this paper, we propose **VIPGuard**, a unified multimodal framework for detecting and explaining deepfakes targeting specific users. VIPGuard addresses identity-aware deepfake detection by explicitly incorporating known facial identity priors, including both global identity information and detailed structural facial attributes from the VIPs. To achieve this, this paper, for the first time, reformulates forgery detection as a *fine-grained face recognition task*, targeted at VIP identities. To leverage both global and local facial information, we use pre-trained face models [62, 14, 11] to extract global facial priors (face similarity scores) and local facial priors (facial attributes). Leveraging these priors, we aim to enable an MLLM to perform forgery detection through semantic comparisons between suspect and authentic faces across visual and textual inputs. We train VIPGuard in three stages: (1) Fine-tuning an MLLM with a variety of facial attribute data to enhance facial understanding; (2) Performing discriminative learning to distinguish some arbitrary identities from manipulated or similar-looking faces by reasoning; (3) Supporting personalized detection by learning a unique and lightweight VIP token to represent each target identity for customized reasoning.

¹In this work, we refer to such individuals as "VIPs" whose identities we aim to protect.

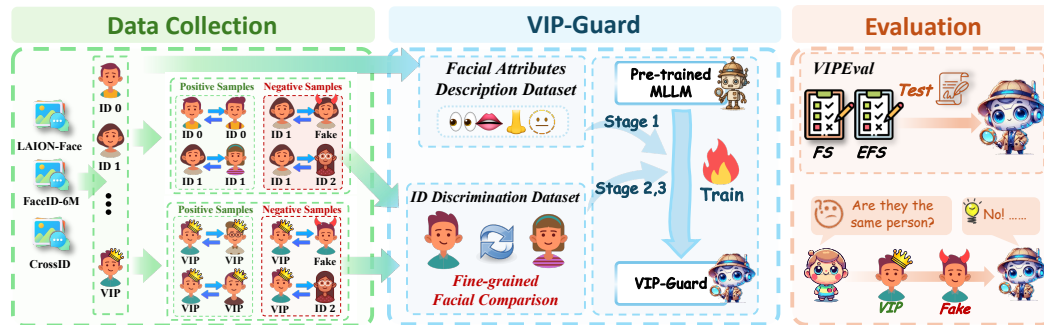


Figure 2: Overview of the data collection, VIP-Guard, and Evaluation. VIPGuard’s training and inference pipeline for facial attribute understanding, identity discrimination, and VIP user customization.

To enable robust evaluation in personalized deepfake detection, we additionally present **VIPBench**—a *comprehensive* and *identity-centric* benchmark that differs fundamentally from conventional deepfake benchmarks [77, 74]. While existing benchmarks typically treat faces generically and overlook whose identity is being manipulated, VIPBench focuses explicitly on identity-aware scenarios, where prior knowledge of the target individual is available. VIPBench includes 22 specific target identities and a total of 80,080 images, covering both real and forged samples. These forgeries are generated using 14 state-of-the-art methods, spanning 7 face-swapping (FS) and 7 entire-face synthesis (EFS) techniques, providing diverse manipulation types and realistic evaluation settings. By centering evaluation around known identities and incorporating fine-grained annotations, VIPBench allows models to be assessed not only on detection accuracy but also on their ability to leverage identity-specific cues.

Our main contributions are summarized as follows:

- We introduce a new formulation for **personalized deepfake detection** that targets specific individuals, casting it as a *fine-grained face recognition problem* based on both global identity features and detailed facial attributes. This formulation requires only a small number of authentic reference images per VIP user, making it practical for real-world personalized protection scenarios.
- We propose **VIPGuard**, a unified multimodal framework for identity-aware deepfake detection and explanation. VIPGuard incorporates pre-trained facial prior models and multimodal large language models, and follows a three-stage pipeline that extracts identity features, performs visual-text reasoning, and supports personalized detection through lightweight identity tokens.
- We introduce **VIPBench**, a comprehensive benchmark for evaluating identity-aware deepfake detection. It consists of **22** real-world target identities and **80,080** images generated by **14** state-of-the-art manipulation methods, enabling fine-grained and realistic assessment of personalized detection performance.

2 Related Works

2.1 General Deepfake Detection

Conventional Deepfake Detection Current deepfake detection faces significant challenges in generalization. To illustrate, researchers have explored a range of approaches, including data augmentation [38, 36, 56, 4, 6], frequency-based cues [50, 43, 88], identity-aware learning [15, 27], disentanglement [78, 76], reconstruction [61, 79], and custom network designs [12]. Data augmentation has proven especially effective for improving generalization—for example, FWA [38] simulates warping artifacts, Face X-ray and SBI [36, 56] target blending boundaries, and SLADD [4] uses adversarial examples to challenge models. Despite these advances, most traditional detectors still offer only binary outputs without human-understandable explanations, leaving users unclear about why a face is classified as fake—limiting trust and explainability.

Deepfake Detection via Multimodal Large Language Model Vision and language are two core modalities in human perception, driving growing interest in visual-language multimodal learning.

In deepfake detection, several studies [31, 55, 80] have explored prompt-based strategies for face forgery analysis, showing that multimodal large language models (MLLMs) offer better explainability than traditional detectors. Others [20, 70, 71, 35, 66, 33, 86, 25] have investigated different MLLMs for explainable detection, while [37] introduced a labeled multimodal dataset to support fine-tuning. Moreover, $\mathcal{X}^2$ -DFD [7] further investigates hybrid frameworks that integrate conventional visual models with MLLMs. However, most of these methods are designed for general-purpose detection and overlook the valuable identity-specific information available in many real-world scenarios.

2.2 Personalized Deepfake Detection for Specific Identity Protection

To detect identity inconsistencies in forged faces, prior works [54, 19, 17, 16, 10, 46] use reference images for personalized deepfake detection. For example, ICT-Ref [16] employs a transformer to detect mismatches between inner- and outer-face regions, while DiffID [81] uses reconstruction-based identity distances to identify fakes. However, these methods mainly rely on global identity features and overlook deeper, user-specific information. This limits their robustness to distribution shifts and prevents them from offering human-understandable explanations.

3 VIPBench: A New Benchmark for Personalized Deepfake Detection

To promote the training and evaluation of personalized deepfake detection, we build a comprehensive identity-aware benchmark called VIPBench for personalized deepfake detection. The training set of VIPBench progressively fine-tunes MLLMs, advancing from basic facial attribute recognition to fine-grained identity inconsistency detection. Moreover, we introduce a new evaluation dataset for identity-aware deepfake detection, a setting that currently lacks sufficient evaluation resources, to assess the effectiveness of different methods. We obtain all facial images from open-sourced datasets, including LAION-Face [87], CrossFaceID [64], and FaceID-6M [63], subjected to some preprocessing (details in supplementary materials). The main idea of the dataset construction pipeline is described below, while complete details are provided in the supplementary materials.

Facial Attributes Description Dataset We present the Facial Attribute Description Dataset—a multimodal dataset composed of high-resolution facial images paired with rich facial attribute descriptions. The dataset ($\mathcal{D}_{FA}$) is collected about 30,000 high-resolution (more than 1024×1024) images from LAION-Face, to facilitate foundational facial understanding in MLLMs. In addition, an MLLM specialized in facial analysis can act as a captioner to generate descriptive facial attribute information. As illustrated in Figure 3, we obtained detailed attributes (e.g., face shape, skin condition) via MegVii’s official API² and subsequently refined them by human experts. Figure 3 (d) provides some examples of these facial attributes, which were then transformed into diverse VQA formats (multiple-choice, short/long answer). The process can be formulated as Eq. 1

$$\mathcal{D}_{FA} = \{(I_i, \text{VQA}(a_{i1}, a_{i2}, \dots, a_{ik})) \mid (a_{i1}, \dots, a_{ik}) \in \mathcal{A}^k, i = 1, \dots, N\}, \quad (1)$$

where $\mathcal{A} = F(I)$ denotes the set of extracted facial attributes, F is the API, and I is the facial image. The dataset $\mathcal{D}_{FA}$ comprises VQA instances generated from all k -tuples of attributes drawn from $\mathcal{A}$. This dataset provides rich supervision for training an MLLM to understand fundamental facial characteristics. The other details about $\mathcal{D}_{FA}$ are shown in the supplementary material.

Identity Discrimination Dataset By reformulating personalized deepfake detection as a target-face-centric and fine-grained face recognition problem, we constructed the Identity Discrimination Dataset $\mathcal{D}_{ID}$. This dataset, comprising $\mathcal{D}_{ID}^{general}$ and $\mathcal{D}_{ID}^{vip}$, includes facial images and corresponding annotations intended for reasoning about fine-grained identity discrimination. Specifically, $\mathcal{D}_{ID}$ comprises facial image pairs: positive (same-identity, real-real) and negative (all others). As shown in the ‘Data Collection’ part of Figure 1, we built facial pairs centered on VIP users in $\mathcal{D}_{ID}^{vip}$, while facial pairs in $\mathcal{D}_{ID}^{general}$ were built using arbitrary identities. Given the high resemblance of forged faces to genuine ones, which is challenge for conventional face recognition, we augmented negative samples using SimSwap [5] (face swapping) and Arc2Face [49] (entire face synthesis). These negative pairs include different-identity real-real pairs and real-versus-forgery pairs, with these two categories balanced roughly 1:1. As demonstrated in Figure 3 (b), for each image pair, we first generated a

²<https://www.faceplusplus.com.cn/>

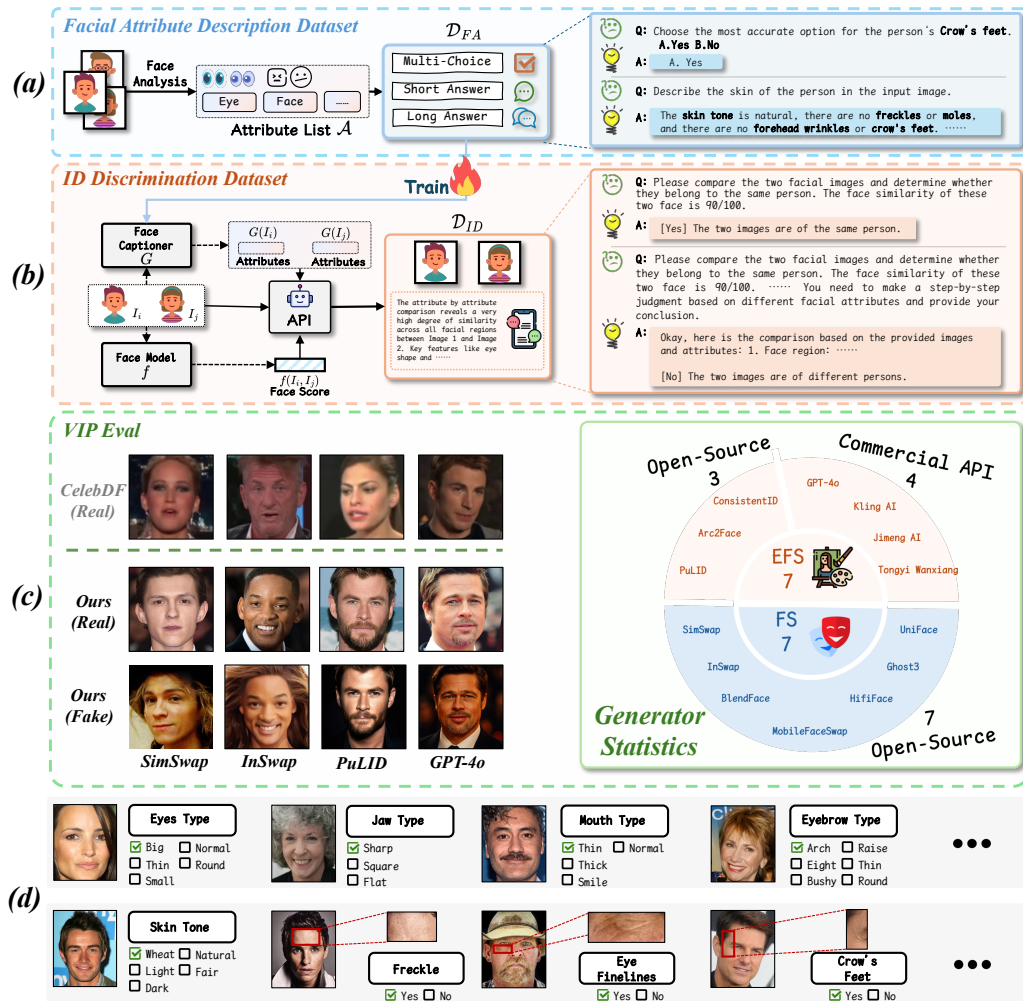


Figure 3: Illustration of the proposed VIPBench, which includes three personalized datasets, (a) Facial Attributes Description Dataset $\mathcal{D}_{FA}$, (b) Identity Discrimination Dataset $\mathcal{D}_{ID}$, and (c) VIPEval $\mathcal{D}_{Eval}$. (d) Some examples of the facial attributes used in the $\mathcal{D}_{FA}$ are also illustrated here, while the full set is available in the supplementary material. The real images shown in (c) are from CelebDF [87] and VIPBench, while the fake ones are generated using multiple models. All images in (d) are sourced from LAION-Face [87].

facial attribute list using our fine-tuned Qwen2.5VL-7B model [1] G , trained on $\mathcal{D}_{FA}$. Then, we acquired the facial similarity scores by a pre-trained facial recognition model f . Finally, we utilized a commercial model (Gemini 2.5 Pro³ [21]) was selected in this paper), denoted as API in Figure 3 (b), to form the training data. Specifically, the API was then strictly required to reason identity discrepancies between image pairs based solely on the provided facial attributes $G(I_i)$ and $G(I_j)$ and similarity scores $f(I_i, I_j)$, ensuring that the analysis was grounded in real features and not influenced by hallucinated content. The above process can be formulated as

$$\begin{aligned} \text{Prompt} &= \text{API}(Q, f(I_i, I_j), I_i, I_j, G(I_i), G(I_j)) \\ \mathcal{D}_{ID} &= \{I_i, I_j, \text{VQA}(I_i, I_j, \text{Prompt}) \mid I_i, I_j \in \mathcal{J}\}, \end{aligned} \quad (2)$$

where I_i and I_j are facial images, Q is the prompt inputted into Gemini, and $\mathcal{J}$ denotes a pre-constructed pool of image-name pairs (see supplemental material for details).

VIPEval (User-Specific Evaluation Dataset): We introduce a user-specific evaluation dataset called **VIPEval** for assessing user-specific forgery detection performance. Conventional datasets [39,

³Gemini API version in use: 2.5-pro-exp-03-25.

[53] typically lack a sufficient number of high-resolution real images per individual, along with corresponding variations across different face forgery methods, which makes it challenging to evaluate personalized forgery detection approaches. The dataset includes images from rich sources and various resolutions with diverse generation methods, making it more representative of real-world conditions. To address this limitation, we carefully selected 22 unique identities from the previously described image-name pool $\mathcal{I}$, ensuring no overlap with the identities in the $\mathcal{D}_{ID}^{general}$. For each identity, 40-60 real images were collected, 20 reserved for testing in this benchmark $\mathcal{D}_{Eval}$, and the remainder used to construct $\mathcal{D}_{ID}^{vip}$. The benchmark includes reserved real images and a comprehensive set of forged counterparts. For each of the 22 test identities, 420 images are generated per method using seven distinct face-swapping (FS) techniques [5, 57, 30, 67, 69, 65, 18]. Additionally, three open-source entire-face synthesis (EFS) methods [49, 28, 24] generate 10 images per real test image by varying random seeds or prompts, resulting in 200 images per identity. Furthermore, four commercial API-based EFS methods [48, 32, 60, 34] produce 20 images per identity. In total, the dataset comprises 80,080 images.

4 VIPGuard: A Multimodal Framework for Personalized Deepfake Detection

4.1 Problem Formulation and Comparison with Prior Works

We consider a personalized deepfake detection scenario in which the detector has access to several authentic images of the target user (e.g., a celebrity, hereafter referred to as a VIP), along with a collection of real facial images from other unrelated individuals. These authentic images serve as prior knowledge for the detector to better recognize and protect the target individual. The objective is to model user-specific facial characteristics to identify suspicious images and protect the VIP's identity from forgery attacks. For forgeries, such as face-swapped images, we impose a realistic constraint that the source identities used for manipulation (denoted as ID^{source}) are unseen by the detector during training. This assumption is closer to the real-world scenario, where attackers can use arbitrary faces that are not available to the defender. This setting contrasts with identity-aware detection methods [16, 19, 46], which assume that both the target identity (ID^{vip}) and source identities (ID^{source}) are known during training—i.e., all relevant faces are included in the reference set $\{ID^{vip}, ID^{source}, ID^{others}\}$. Such a closed-world assumption simplifies the problem but is rarely realistic in practical applications. In our formulation, the detector only has access to $\{ID^{vip}, ID^{others}\}$ during training, while ID^{source} remains unknown. This difference introduces a more challenging yet practical problem setting, emphasizing the need for generalization to unseen source identities and better alignment with real-world deployment scenarios.

4.2 Method Overview

To address the challenge, we propose a framework, VIPGuard, which develops an MLLM capable of identifying suspicious images of a specific user, while providing human-understandable explanations based on the user's unique facial attributes. We reformulate personalized deepfake detection as a fine-grained face recognition problem centered on the protected target, where forgeries are detected through the MLLM's reasoning over global identity features and detailed facial attributes. To equip the model with this capability, we start from a pre-trained MLLM (Qwen-2.5-VL-7B [1]) and progressively fine-tune it through a three-stage process: Face Attribute Learning, Identity Discrimination, and User-Specific Customization. We describe each stage in detail below.

4.3 Three-Stages Training of VIPGuard

Stage 1: Face Attributes Learning It is crucial first to enhance the model's capability to recognize and utilize fine-grained facial attributes, as naive MLLMs inherently lack a sufficient understanding of human facial features for effective VIP identity protection. To this end, we fine-tune the pre-trained MLLM on $\mathcal{D}_{FA}$, which contains a large number of samples for facial attribute recognition (see Section 3). We integrate LoRA [26] modules into the pre-trained MLLM and perform supervised fine-tuning using an autoregressive loss, as defined in Eq. 3:

$$L(\theta) = - \sum_{i=1}^N \log [p_{\theta}(x_i | x_{<i}, E_V(I))], \quad (3)$$

where θ denotes the parameters of both the inserted LoRA modules and the original MLLM, N is the length of the output prompt, x_i is the i -th token to be predicted while $x_{<i}$ are the previous tokens,

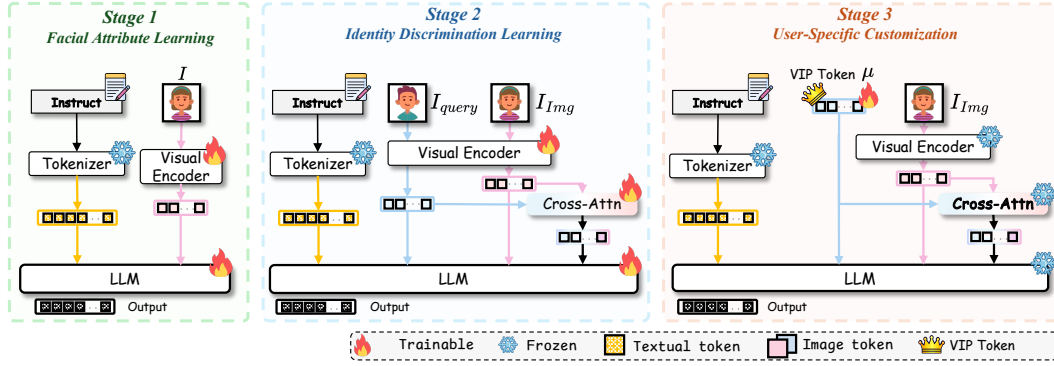


Figure 4: Illustration of the three stages of training the proposed VIPGuard framework.

E_V represents the visual encoder of the MLLM, and I is the input image. After training, the MLLM can understand and then recognize facial attributes.

Stage 2: Identity Discrimination Learning As illustrated in Figure 1, although the fake face appears realistic at first glance, subtle discrepancies remain in its detailed facial attributes. Therefore, we reformulate personalized deepfake detection as a fine-grained face recognition problem centered on the protected target, where forgeries are detected through the MLLM’s reasoning over global identity features and detailed facial attributes. In this stage, we further fine-tune the MLLM on $\mathcal{D}_{ID}^{general}$ for fine-grained facial recognition across face pairs—each consisting of an arbitrary query face I_{query} and an arbitrary input face I_{img} —thereby equipping it with a foundational ability to identity reason. $\mathcal{D}_{ID}^{general}$ contains a large number of positive (same-identity) and negative (including both different-identity and forgery) face pairs, each annotated with VQA-style reasoning questions and answers. These annotations support fine-grained identity discrimination by reasoning over both global and local facial prior. Here, we employ face recognition models [62, 14, 11] to provide face similarity scores, serving as a global facial prior, while the local facial prior—i.e., knowledge of facial attributes—has already been incorporated into the fine-tuned MLLM during Stage 1. Specifically, as shown in Figure 4, the query I_{query} and the input image I_{img} are first inputted into the visual encoder E_V of our MLLM to obtain the visual token sequences f_{query} and f_{img} . Pre-trained MLLMs typically lack task-specific optimization for facial recognition, rendering the vanilla visual features f_{query} and f_{img} suboptimal for facial identity discrimination. Hence, we introduce a Cross-Attention (Cross-Attn) module [40] to capture fine-grained differences in visual tokens between distinct facial images, which is formulated as follows

$$g = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_K}}\right)V, \quad \text{where } Q = f_{query}, K = V = f_{img}. \quad (4)$$

We optimize the MLLM by an autoregressive loss shown as Eq. 5 below.

$$L(\theta) = -\sum_{i=1}^N \log [p_{\theta}(x_i \mid x_{<i}, f_{query}, f_{img}, g)]. \quad (5)$$

After training, the model can perform fine-grained facial recognition between any two faces, while also providing detailed explanations of their differences.

Stage 3: User-Specific Customization Upon completing Stage 2, the MLLM acquires a basic capability for fine-grained facial identity comparison with reasoning. In Stage 3, we further introduce a user-specific customization to enable personalized forgery detection for a given VIP user. To this end, we construct a dataset $\mathcal{D}_{ID}^{vip}$ with the same structure as $\mathcal{D}_{ID}^{general}$, centered on the VIP’s identity, allowing the model to identify suspicious images by reasoning based on the VIP user’s facial prior. Motivated by the Yo’LLaVA [47], to facilitate lightweight deployment, we incorporate a learnable VIP token trained on $\mathcal{D}_{ID}^{vip}$, denoted as μ , which encodes identity-specific features of the VIP user. During this stage, the parameters of the MLLM from Stage 2 are frozen, and only the VIP token is trained on $\mathcal{D}_{ID}^{vip}$, thereby refining the model’s ability to perceive and distinguish the target VIP user. As illustrated in Figure 4, the learnable VIP token μ , which is a vector of size $32 \times d$, is employed to

substitute the visual feature representation f_{query} of the query image I_{query} . Formally, the prediction procedure can be described as follows

$$g = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_K}}\right) V, \quad \text{where } Q = \mu, K = V = f_{\text{img}}. \quad (6)$$

$$L(\mu) = - \sum_{i=1}^N \log[p_\theta(x_i | x_{<i}, \mu, f_{query}, g)]. \quad (7)$$

Notably, the query images in $\mathcal{D}_{ID}^{vip}$ are not used; only the input image I_{img} and the corresponding reasoning annotation are fed into the MLLM. After training, the MLLM can perform personalized forgery detection for the VIP user by determining whether the identity of the input image I_{img} belongs to the target individual and providing a user-centric explanation. Furthermore, during inference, we can substitute the corresponding VIP token μ for each user, enabling efficient and accurate detection of malicious facial forgeries without requiring any model retraining. This design supports lightweight and scalable deployment.

5 Experiments

In this section, we present comprehensive experiments to evaluate the effectiveness of our method. For general deepfake detectors, we used models trained on standard deepfake datasets. For ID-aware detectors, we assumed access to 20-30 real images of each VIP user. Detailed experimental settings are provided in the supplementary material.

Evaluation Metrics For evaluation, we report three standard metrics: Area Under the Curve (AUC), Equal Error Rate (EER), and Accuracy (ACC). AUC measures the model’s ability to distinguish between positive and negative classes across all thresholds. EER represents the point where the false acceptance rate is equal to the false rejection rate. ACC denotes the proportion of correct predictions. We use AUC and EER to compare our method following prior work [74, 77]. Moreover, ACC is used to compare with API-based commercial generators (e.g., GPT-4o⁴ [48]).

Comparison with Deepfake Detection Methods on VIFEval We evaluated general deepfake detectors, ID-aware detectors, and our proposed method on VIFEval, reporting results for face swapping (Table 1) detection and entire faces synthesis detection (Table 2). Our method significantly improves the detection of all facial forgery types. General deepfake detectors (e.g., Effort [73]) excel at face swapping detection but fail to detect more realistic forgeries from commercial APIs due to unseen artifacts. ID-aware detectors, unlike general detectors reliant on low-level cues, leverage identity-related semantic consistency for robust detection across diverse generation techniques and data. However, existing ID-aware methods (e.g., DiffID, ICT-Ref) struggle with fully synthesized faces due to their reliance on global facial features, a limitation amplified by advancing forgery techniques. Conversely, VIP-Guard captures identity-specific cues by jointly leveraging global facial representations and local attribute analysis, enabling robust protection of VIP users with few real photos across diverse forgery methods.

Table 1: Evaluation of generalization performance (AUC (%) / EER (%)) for face swapping detection on VIFEval.

Methods	BlendFace [57]	Ghost3 [18]	HifiFace [65]	InSwap [30]	MobileSwap [69]	SimSwap [5]	UniFace [67]
Xception [9]	53.89 / 46.59	61.08 / 42.27	71.70 / 34.09	64.79 / 38.64	98.12 / 7.50	64.91 / 38.41	59.91 / 42.50
EfficientNet [58]	33.94 / 60.91	43.52 / 54.32	61.22 / 42.95	37.34 / 58.64	89.67 / 64.45	56.55 / 45.00	46.53 / 50.23
UCF [76]	64.31 / 38.18	65.92 / 35.68	65.62 / 36.82	66.02 / 34.32	87.33 / 20.00	63.11 / 36.82	67.25 / 34.32
ProDet [8]	59.84 / 42.95	53.18 / 47.73	56.48 / 46.36	35.91 / 59.32	73.01 / 34.32	56.27 / 45.68	49.34 / 50.00
RECCE [3]	56.11 / 45.00	61.53 / 39.54	56.93 / 45.91	58.85 / 43.41	88.40 / 19.32	60.05 / 42.95	63.01 / 39.10
CDFA [42]	59.86 / 45.23	70.50 / 36.82	85.43 / 24.09	70.92 / 36.59	98.75 / 5.00	71.22 / 35.91	73.30 / 34.77
RepDFD [41]	70.34 / 35.00	78.76 / 28.18	80.09 / 27.50	71.66 / 34.32	95.76 / 10.45	79.40 / 27.95	82.36 / 25.45
Effort [73]	91.87 / 17.61	95.28 / 13.07	96.60 / 10.23	85.53 / 23.86	96.23 / 13.07	97.16 / 8.52	92.48 / 13.64
DiffID [81]	83.33 / 24.47	66.02 / 38.38	87.23 / 20.33	66.22 / 38.06	75.71 / 30.89	72.03 / 33.91	83.30 / 24.40
ICT-Ref [16]	88.67 / 13.37	86.52 / 15.51	85.90 / 14.05	84.34 / 17.87	87.45 / 13.52	86.57 / 14.65	82.73 / 19.04
VIPGuard	99.48 / 0.81	97.97 / 4.39	99.63 / 0.61	96.40 / 8.24	99.55 / 1.01	99.43 / 1.02	99.69 / 0.26

⁴GPT4o API version in use: GPT-4o-2024-08-06

Table 2: Evaluation of generalization performance (AUC (%) / EER (%)) for entire face synthesis detection on VIPEval.

Method	Open-Source			Commercial-API			
	ConsistentID [28]	Arc2Face [49]	PuLID [24]	GPT-4o [48]	Jimeng AI [32]	TongYi [60]	Kling AI [34]
Xception	42.02 / 54.77	51.87 / 48.86	59.23 / 44.09	58.13 / 46.36	57.77 / 44.55	34.36 / 62.05	44.34 / 54.32
EfficientNet	33.81 / 61.14	44.96 / 50.91	47.19 / 50.23	75.35 / 28.86	55.04 / 45.00	41.15 / 55.68	42.40 / 54.32
UCF [76]	62.16 / 40.91	56.62 / 46.36	54.16 / 46.36	59.06 / 42.73	71.08 / 32.78	82.38 / 22.73	63.31 / 40.23
ProDet [8]	63.68 / 37.95	59.62 / 40.91	67.02 / 36.82	59.23 / 40.91	59.71 / 42.73	89.80 / 18.41	72.53 / 32.73
RECCE [3]	68.56 / 36.82	57.79 / 47.72	63.85 / 40.00	83.63 / 22.73	69.00 / 35.23	97.00 / 7.50	70.94 / 37.72
CDFA [42]	77.62 / 30.00	67.09 / 39.09	67.93 / 37.50	73.46 / 32.73	71.98 / 34.09	90.32 / 17.95	77.47 / 30.00
RepDFD [41]	83.52 / 24.32	61.67 / 41.59	74.65 / 32.27	73.62 / 32.27	62.78 / 40.91	93.80 / 14.09	60.10 / 43.41
Effort [73]	58.68 / 47.35	57.03 / 44.89	56.31 / 44.89	49.63 / 48.30	63.60 / 40.34	82.93 / 23.86	56.84 / 44.89
DiffID [81]	75.85 / 32.01	78.47 / 28.86	70.36 / 35.72	45.26 / 52.97	64.29 / 39.35	84.66 / 23.58	69.51 / 35.56
ICT-Ref [16]	63.15 / 39.84	70.27 / 32.84	72.36 / 31.93	58.59 / 40.93	65.36 / 35.73	74.88 / 24.41	50.05 / 45.98
VIPGuard	99.69 / 0.45	98.05 / 4.80	98.96 / 1.90	89.03 / 16.14	97.04 / 5.36	99.76 / 0.23	99.27 / 1.25

Comparison with Other LLM-based Methods on VIPEval This experiment evaluated the detection capabilities of Multimodal Large Language Models (MLLMs) on VIPEval by comparing various MLLM-based methods, including FFAA [29], an MLLM specialized for face forgery detection, alongside naive MLLMs. Detection performance was measured using Accuracy (ACC) due to the API’s binary (Real/Fake) output. As presented in Table 3, our method consistently outperforms other approaches across all forged image types, demonstrating the effectiveness of identity-specific semantic detection.

Table 3: Comparison (ACC (%)) of our method and other MLLMs on the VIPEval.

Method	BlendFace	HifiFace	MobileSwap	UniFace	ConsistentID	Arc2Face
GPT-4o-2024-08-06[48]	71.85	84.36	94.35	70.72	46.70	50.90
Gemini-2.5-pro-exp-03-25 [21]	80.87	82.68	91.75	83.58	69.85	79.98
Qwen2.5VL 7B [1]	50.05	49.86	49.99	50.02	49.51	49.69
LLaMA3.2Vision 11B [22]	44.76	43.93	45.71	54.09	38.56	43.61
FFAA [29]	88.60	89.95	91.36	89.30	63.53	59.89
VIPGuard	95.51	95.82	96.71	95.88	95.91	89.71

One-shot Performance on other Deepfake detection datasets This experiment evaluates our method against several existing ID-aware approaches on established benchmark datasets. The evaluation specifically employed the challenging CelebDF [39] dataset and a CelebDF-related subset of DF40 [74]. Owing to limited real sample diversity in these datasets, our method was evaluated in a one-shot setting, using a single real image per user. To mitigate similarity from shared video sources, reference and test images were sampled from different videos. As shown in Table 4, our method still achieves competitive performance compared to other approaches. Notably, due to limited real image availability, these results were obtained using VIP-Guard at Stage 2, omitting Stage 3. The results also verify the effectiveness of Stage 2 of our method.

Table 4: Evaluation of frame-level performance (AUC (%) / EER (%)) in CelebDF [39] and DF40 [74] under one-shot setting. For each identity, only a single real image is available.

Method	CelebDF	DF40						
		BlendFace	E4S	FaceDancer	FSGAN	SimSwap	UniFace	InSwap
Diff-ID	86.66 / 21.79	76.83 / 28.99	84.79 / 24.07	81.04 / 27.02	82.88 / 23.77	64.87 / 38.28	84.04 / 22.53	59.62 / 43.19
ICT-Ref	81.86 / 26.18	76.47 / 30.63	83.33 / 24.93	91.57 / 16.38	73.62 / 32.95	82.41 / 25.62	80.34 / 27.07	69.91 / 35.98
VIPGuard	87.96 / 20.84	81.87 / 26.23	89.74 / 19.63	86.12 / 23.61	86.44 / 20.87	83.65 / 24.92	86.70 / 22.57	70.97 / 34.15

Visual Examples of Our Model Explanations Figure 5 presents two representative examples illustrating VIPGuard’s ability to detect anomalous local facial attributes in deepfake images generated by entire face synthesis (EFS, left) and face swapping (FS, right) techniques. Despite the high global facial similarity between the real and fake images, VIPGuard accurately identifies subtle but critical differences in localized regions such as the eyes, lips, and skin textures. For instance, discrepancies in eye size, lip thickness, and specific skin features (e.g., glabella wrinkles, eye pouches, moles, and crow’s feet) are effectively captured by our model. These examples highlight two key strengths of VIPGuard: (1) its ability to detect forgery-induced anomalies in local facial regions where current

generation methods often fail, and (2) its capacity to reason over these attributes for reliable identity verification, even when global facial appearance is highly similar.

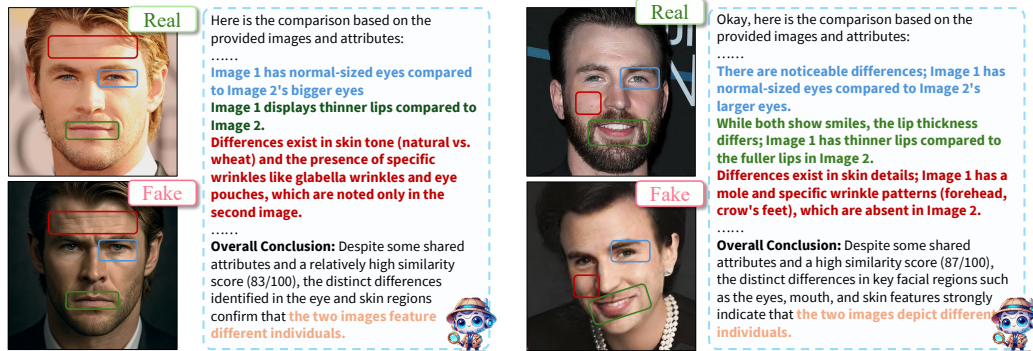


Figure 5: Visual illustration of the analysis of VIP-Guard detecting anomalous local facial attributes for EFS (left) and FS (right). The two real images are sourced from LAION-Face [87], while the fake images in the left and right subfigures were generated by GPT-4o [48] and HifiFace [65], respectively.

5.1 Effectiveness in Annotation-Free Scenarios

To validate VIPGuard’s practicality, we tested its Stage 3 performance when trained using only VIP images, foregoing textual annotations. As shown in Table 5, the model’s performance remains strong, with the average AUC decreasing only slightly from 98.98% to 95.86%. This robustness stems directly from the design of our framework. Stage 2 is responsible for internalizing a vast repository of discriminative knowledge by learning to distinguish between arbitrary image pairs with the aid of textual supervision. This foundational pre-training is so effective that the User-specific fine-tuning in Stage 3 requires only visual information to achieve a high degree of accuracy. Consequently, this experiment validates VIPGuard as a highly practical and readily deployable framework for real-world applications, including annotation-free scenarios. The model’s ability to achieve such high efficacy under these constraints is a direct testament to the comprehensive and foundational knowledge provided by Stage 2.

Table 5: Performance (AUC (%)) of VIPGuard under different training configurations in Stage 3. *Images + Annotation* denotes training with both images and textual descriptions, while *Only Images* uses visual inputs only.

Variant	BlendFace	InSwap	Arc2Face	PuLID	Average
<i>Only Images</i>	98.45	92.91	94.35	97.72	95.86
<i>Images + Annotation</i>	99.48	99.43	98.05	98.96	98.98

6 Conclusion

This paper proposes VIPGuard, addressing a critical gap in deepfake detection by leveraging known facial identities to enable personalized, accurate, and explainable detection. Unlike traditional detectors that mainly rely on low-level visual artifacts or general-purpose MLLMs lacking identity awareness, VIPGuard integrates fine-grained attribute learning, identity-level discriminative training, and user-specific customization through a unified multimodal framework. Combined with our newly proposed VIPBench, which enables rigorous evaluation across diverse and advanced forgery types, our approach demonstrates clear superiority in both detection performance and explainability, offering a robust and scalable solution for safeguarding high-risk individuals against identity-based deepfakes.

Content Structure of the Appendix. Due to page constraints, we include additional analyses and experiments in the Appendix, containing comprehensive ablation studies (Appendix D.2), robustness evaluation (Appendix D.4), adaptive choice of VIP Token (Appendix D.3), more visual examples of model explanations (Appendix D.5), details of dataset construction (Appendix E), an ethical statement (Appendix A), and limitations and future work (Appendix C). *For further details, please refer to the Appendix.*

Acknowledgments and Disclosure of Funding

Dr Bin Li was supported in part by NSFC and in part by Shenzhen R&D Program (Grant JCYJ20250604181211016, SYSPG20241211174032004). Dr Weixiang Li was supported in part by NSFC (Grant 62202310, 62572328) and in part by Guangdong Basic and Applied Basic Research Foundation (Grant 2025A1515010292).

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [2] Matyas Bohacek and Hany Farid. Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms. *Proceedings of the National Academy of Sciences*, 119(48):e2216035119, 2022.
- [3] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022.
- [4] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18710–18719, 2022.
- [5] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2003–2011, 2020.
- [6] Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, et al. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025.
- [7] Yize Chen, Zhiyuan Yan, Siwei Lyu, and Baoyuan Wu. x^2 -dfd: A framework for explainable and extendable deepfake detection. *arXiv preprint arXiv:2410.06126*, 2024.
- [8] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? *arXiv preprint arXiv:2408.17052*, 2024.
- [9] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.
- [10] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15108–15117, 2021.
- [11] Jun Dan, Yang Liu, Haoyu Xie, Jiankang Deng, Haoran Xie, Xuansong Xie, and Baigui Sun. Transface: Calibrating transformer training for face recognition from a data-centric perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20642–20653, October 2023.
- [12] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [13] DeepFaceLab. <https://github.com/iperov/DeepFaceLab>.
- [14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

- [15] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3994–4004, 2023.
- [16] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9468–9478, June 2022.
- [17] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Identity-driven deepfake detection. *arXiv preprint arXiv:2012.03930*, 2020.
- [18] FaceFusion 3.0. <https://learn.rundiffusion.com/facefusion-3-0/>, 2025.
- [19] Mingqi Fang, Lingyun Yu, Hongtao Xie, Qingfeng Tan, Zhiyuan Tan, Amir Hussain, Zezheng Wang, Jiahong Li, and Zhihong Tian. Stidnet: Identity-aware face forgery detection with spatiotemporal knowledge distillation. *IEEE Transactions on Computational Social Systems*, 11(4):5354–5366, 2024.
- [20] Niki M Foteinopoulou, Enjie Ghorbel, and Djamila Aouada. A hitchhiker’s guide to fine-grained face forgery detection using common sense reasoning. *Advances in Neural Information Processing Systems*, 37:2943–2976, 2025.
- [21] Google DeepMind. Gemini 2.5: Our most intelligent ai model. <https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/>, 2025.
- [22] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [23] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z. Li. Towards fast, accurate and stable 3d dense face alignment. In *Computer Vision – ECCV 2020*, pages 152–168, 2020.
- [24] Zinan Guo, Yanze Wu, Chen Zhuowei, Peng Zhang, Qian He, et al. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in neural information processing systems*, 37:36777–36804, 2024.
- [25] Xinan He, Yue Zhou, Bing Fan, Bin Li, Guopu Zhu, and Feng Ding. Vlforgery face triad: Detection, localization and attribution via multimodal large language models. *arXiv preprint arXiv:2503.06142*, 2025.
- [26] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- [27] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiabin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4490–4499, 2023.
- [28] Jiehui Huang, Xiao Dong, Wenhui Song, Zheng Chong, Zhenchao Tang, Jun Zhou, Yuhao Cheng, Long Chen, Hanhui Li, Yiqiang Yan, et al. Consistentid: Portrait generation with multimodal fine-grained identity preserving. *arXiv preprint arXiv:2404.16771*, 2024.
- [29] Zhengchao Huang, Bin Xia, Zicheng Lin, Zhun Mou, and Wenming Yang. Ffaa: Multimodal large language model based explainable open-world face forgery analysis assistant. *arXiv preprint arXiv:2408.10072*, 2024.
- [30] inswapper. <https://github.com/haofanwang/inswapper>.

- [31] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4324–4333, 2024.
- [32] JiMeng-AI. <https://jimeng-ai.org/>, 2025.
- [33] Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, et al. Legion: Learning to ground and explain for synthetic image detection. *arXiv preprint arXiv:2503.15264*, 2025.
- [34] Kling AI. <https://www.klingai.com/global/>, 2025.
- [35] Jiawei Li, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. *arXiv preprint arXiv:2410.10238*, 2024.
- [36] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [37] Yixuan Li, Xuelin Liu, Xiaoyang Wang, Shiqi Wang, and Weisi Lin. Fakebench: Uncover the achilles’ heels of fake images with large multimodal models. *arXiv preprint arXiv:2404.13306*, 2024.
- [38] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.
- [39] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A new dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [40] Hezheng Lin, Xing Cheng, Xiangyu Wu, and Dong Shen. Cat: Cross attention in vision transformer. In *2022 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2022.
- [41] Kaiqing Lin, Yuzhen Lin, Weixiang Li, Taiping Yao, and Bin Li. Standing on the shoulders of giants: Reprogramming visual-language model for general deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5262–5270, 2025.
- [42] Yuzhen Lin, Wentang Song, Bin Li, Yuezun Li, Jiangqun Ni, Han Chen, and Qiushi Li. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. *arXiv preprint arXiv:2409.14444*, 2024.
- [43] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.
- [44] Long Ma, Zhiyuan Yan, Yize Chen, Jin Xu, Qinglang Guo, Hu Huang, Yong Liao, and Hui Lin. From specificity to generality: Revisiting generalizable artifacts in detecting face deepfakes. *arXiv preprint arXiv:2504.04827*, 2025.
- [45] Meta AI. Introducing llama 3.2: Vision, edge, and mobile models. <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>, 2024.
- [46] Liyue Ming, Peisong He, Haoliang Li, Shiqi Wang, and Xinghao Jiang. Critical contour prior-guided graph learning with pose calibration for identity-aware deepfake detection. *IEEE Transactions on Multimedia*, 2025.
- [47] Thao Nguyen, Haotian Liu, Yuheng Li, Mu Cai, Utkarsh Ojha, and Yong Jae Lee. Yo’llava: Your personalized language and vision assistant. *arXiv preprint arXiv:2406.09400*, 2024.
- [48] OpenAI. GPT-4o. <https://openai.com/index/hello-gpt-4o>, 2025.

- [49] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *European Conference on Computer Vision*, pages 241–261. Springer, 2024.
- [50] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference Computer Vision*, pages 86–103. Springer, 2020.
- [51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [52] Roop. <https://github.com/s0md3v/roop>.
- [53] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019.
- [54] Dongyao Shen, Youjian Zhao, and Chengbin Quan. Identity-referenced deepfake detection with contrastive learning. In *Proceedings of the 2022 ACM Workshop on Information Hiding and Multimedia Security*, pages 27–32, 2022.
- [55] Yichen Shi, Yuhao Gao, Yingxin Lai, Hongyang Wang, Jun Feng, Lei He, Jun Wan, Changsheng Chen, Zitong Yu, and Xiaochun Cao. Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. *arXiv preprint arXiv:2402.04178*, 2024.
- [56] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022.
- [57] Kaede Shiohara, Xingchao Yang, and Takafumi Taketomi. Blendface: Re-designing identity encoders for face-swapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7634–7644, 2023.
- [58] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.
- [59] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024.
- [60] TongYiWanXiang. <https://tongyi.aliyun.com/wanxiang/>, 2025.
- [61] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14923–14932, 2021.
- [62] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.
- [63] Shuhe Wang, Xiaoya Li, Jiwei Li, Guoyin Wang, Xiaofei Sun, Bob Zhu, Han Qiu, Mo Yu, Shengjie Shen, Tianwei Zhang, et al. Faceid-6m: A large-scale, open-source faceid customization dataset. *arXiv preprint arXiv:2503.07091*, 2025.
- [64] Shuhe Wang, Xiaoya Li, Xiaofei Sun, Guoyin Wang, Tianwei Zhang, Jiwei Li, and Eduard Hovy. Turn that frown upside down: Faceid customization via cross-training data. *arXiv preprint arXiv:2501.15407*, 2025.
- [65] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hiface: 3d shape and semantic prior guided high fidelity face swapping. *arXiv preprint arXiv:2106.09965*, 2021.

- [66] Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *arXiv preprint arXiv:2503.14905*, 2025.
- [67] Chao Xu, Jiangning Zhang, Yue Han, Guanzhong Tian, Xianfang Zeng, Ying Tai, Yabiao Wang, Chengjie Wang, and Yong Liu. Designing one unified framework for high-fidelity face reenactment and swapping. In *European conference on computer vision*, pages 54–71. Springer, 2022.
- [68] Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *arXiv preprint arXiv:2404.10667*, 2024.
- [69] Zhiliang Xu, Zhibin Hong, Changxing Ding, Zhen Zhu, Junyu Han, Jingtuo Liu, and Errui Ding. Mobilefaceswap: A lightweight framework for video face swapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2973–2981, 2022.
- [70] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [71] Zhipei Xu, Xuanyu Zhang, Xing Zhou, and Jian Zhang. Avatarshield: Visual reinforcement learning for human-centric video forgery detection. *arXiv preprint arXiv:2505.15173*, 2025.
- [72] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8994, 2024.
- [73] Zhiyuan Yan, Jiangming Wang, Zhendong Wang, Peng Jin, Ke-Yue Zhang, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024.
- [74] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*, 2024.
- [75] Zhiyuan Yan, Junyan Ye, Weijia Li, Zilong Huang, Shenghai Yuan, Xiangyang He, Kaiqing Lin, Jun He, Conghui He, and Li Yuan. Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. *arXiv preprint arXiv:2504.02782*, 2025.
- [76] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. *arXiv preprint arXiv:2304.13949*, 2023.
- [77] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 4534–4565. Curran Associates, Inc., 2023.
- [78] Tianyun Yang, Juan Cao, Qiang Sheng, Lei Li, Jiaqi Ji, Xirong Li, and Sheng Tang. Learning to disentangle gan fingerprint for fake image attribution. *arXiv preprint arXiv:2106.08749*, 2021.
- [79] Zheng Yang, Ruoxin Chen, Zhiyuan Yan, Ke-Yue Zhang, Xinghe Fu, Shuang Wu, Xiujun Shu, Taiping Yao, Shouhong Ding, and Xi Li. All patches matter, more patches better: Enhance ai-generated image detection via panoptic patch learning. *arXiv preprint arXiv:2504.01396*, 2025.
- [80] Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, et al. Loki: A comprehensive synthetic data detection benchmark using large multimodal models. *arXiv preprint arXiv:2410.09732*, 2024.
- [81] Chuer Yu, Xuhong Zhang, Yuxuan Duan, Senbo Yan, Zonghui Wang, Yang Xiang, Shouling Ji, and Wenzhi Chen. Diff-id: An explainable identity difference quantification framework for deepfake detection. *IEEE Transactions on Dependable and Secure Computing*, 21(5):5029–5045, 2024.

- [82] Zeqin Yu, Bin Li, Yuzhen Lin, Jinhua Zeng, and Jishen Zeng. Learning to locate the text forgery in smartphone screenshots. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [83] Zeqin Yu, Jiangqun Ni, Yuzhen Lin, Haoyi Deng, and Bin Li. Diffforensics: Leveraging diffusion prior to image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12765–12774, 2024.
- [84] Zeqin Yu, Jiangqun Ni, Jian Zhang, Haoyi Deng, and Yuzhen Lin. Reinforced multi-teacher knowledge distillation for efficient general image forgery detection and localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 995–1003, 2025.
- [85] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE SPL*, 23(10):1499–1503, 2016.
- [86] Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *European Conference on Computer Vision*, pages 399–415. Springer, 2024.
- [87] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022.
- [88] Yue Zhou, Bing Fan, Pradeep K. Atrey, and Feng Ding. Exposing deepfakes using dual-channel network with multi-axis attention and frequency analysis. In *Proceedings of the 2023 ACM Workshop on Information Hiding and Multimedia Security*, pages 169–174, 2023.
- [89] Yue Zhou, Xinan He, KaiQing Lin, Bin Fan, Feng Ding, and Bin Li. Breaking latent prior bias in detectors for generalizable aigc image detection. *arXiv preprint arXiv:2506.00874*, 2025.

Appendix

A Ethics Statement

This study involves facial forgery detection and may raise privacy and ethical concerns related to the use of human facial data. We affirm that all source images used in our datasets are obtained from publicly available and legally compliant open-source datasets [87, 63, 64], which are intended for academic research purposes. Our research is dedicated to advancing facial privacy protection and deepfake detection. The developed models and datasets are intended solely for ethical, academic use and will not be released for commercial or malicious purposes. This study follows the ethical guidelines provided by NeurIPS Code of Ethics, and does not involve any personally identifiable information collected directly by the authors. Our VIPBench dataset is released under the Creative Commons Attribution-NonCommercial (CC BY-NC) license (more details can be seen in <https://creativecommons.org/licenses/by-nc/4.0/>). Furthermore, access to the dataset will be managed through a request form (hosted on HuggingFace) to monitor and control its usage. All interested parties are required to complete the form, and each request will be manually reviewed to help prevent potential misuse.

B Boarder Impact

The research presented in this paper introduces VIPGuard, a multimodal framework for personalized deepfake detection, and VIPBench, a benchmark dataset for evaluating identity-aware detectors. This work addresses a critical and growing societal challenge: the malicious misuse of AI-generated media, particularly in the form of deepfakes targeting specific individuals such as celebrities. Our method promotes positive societal impact in several key areas:

- **Mitigating Disinformation and Identity Theft:** Deepfakes pose a significant threat by enabling the creation of highly realistic fake media, which can be weaponized to harm reputations, manipulate public perception, or conduct fraud. VIPGuard offers a targeted defense mechanism, enhancing the security and privacy of individuals by providing personalized and explainable detection tools.
- **Promoting Safer AI Ecosystems:** By releasing VIPBench, we aim to catalyze further research in robust, personalized detection. Public benchmarks encourage accountability, replication, and the development of defenses that are grounded in real-world threats. Furthermore, we will implement a review-based distribution process for the dataset to ensure its lawful and responsible use.

C Limitation and Future Works

VIPGuard is designed to protect the identities of target users against deepfakes. While the method demonstrates strong generalization capabilities, it does not yet fully exploit additional modalities such as audio, 3D facial models, temporal consistency, and other complementary cues. For instance, these modalities could significantly improve the detection of audio-driven manipulation techniques, which rely heavily on temporal dynamics and inter-frame consistency. Beyond facial images, incorporating such prior information could further enhance detection performance. As part of future work, we plan to integrate these complementary modalities to strengthen VIPGuard's effectiveness. In addition, we recognize two key limitations observed during evaluation: (1) **Significant variations** in head pitch and yaw angles, such as noticeable rotations exceeding 50°, can affect facial features and identity information. However, this limitation is shared by most face-based systems, including standard face recognition models. (2) **Significant age differences** between training and test data can lead to discrepancies in performance. For instance, training may rely on younger facial images while testing may involve older ones. This issue, arising from identity shifts across age, can be alleviated by expanding and diversifying the dataset. Inspired by these observations, we plan to address these potential failure cases in future work by incorporating 3D facial information, learning identity-related temporal cues from videos, and exploring other complementary sources of information.

Table 6: An ablation study on the effectiveness (AUC (%)) of different components in VIPGuard

Variant	Description	BlendFace	InSwap	Arc2Face	PuLID
Baseline	Qwen-2.5-VL-7B	51.00	49.79	49.85	50.30
+ Stage 3	Only performing User-Specific Customization	71.96	59.10	62.12	68.18
+ Stage 1, 3	Enhancing facial understanding and then performing User-Specific Customization	96.50	87.38	88.95	94.11
+ Stage 1, 2	Enhancing facial understanding and general identity discrimination between arbitrary face pairs	90.82	73.44	77.25	71.08
+ Stage 1, 2, 3	Enhancing facial understanding, general identity discrimination, and User-Specific Customization	99.48	96.40	98.05	98.96

D Experiments

D.1 Implementation Details

This paper adopted the pre-trained Qwen-2.5-VL-7B model [1] as the backbone, adhering to its default pre-processing settings. Input images were resized to 448×448 when larger than this size. The model was optimized using the Adam optimizer with a cosine learning rate decay schedule, starting from an initial learning rate of $3e-5$. To accommodate GPU memory limitations, the equivalent batch size was maintained at 72 for both Stage 1 and Stage 2 by applying gradient accumulation. In Stage 3, the effective batch size was reduced to 8 and the initial learning rate was set to 1. All training was performed using mixed-precision computation within the open-source Swift⁵ framework. The model was trained for 2 epochs in Stage 1, and for 1 epoch each in Stage 2 and Stage 3. For the evaluation in the main paper, we obtained the AUC and EER metrics by computing the normalized probability of output logits in MLLM’s prediction head for the words ‘Yes’ and ‘No’. During evaluation, VIPGuard was only required to output the final prediction (i.e., ‘Yes’ or ‘No’) without providing any explanatory content.

D.2 Ablation Study

In this section, we conduct a series of ablation studies to comprehensively evaluate the effectiveness of different components and design choices in our proposed VIPGuard framework. The AUC (%) metric was selected to exhibit the performance in detecting multiple facial forgeries.

Exploration for the Composition in VIPGuard In this paper, we propose a three-stage learning framework VIPGuard to respectively improve MLLM’s face understanding, fine-grained discrimination between arbitrary face pairs, and fine-grained discrimination for VIPs. To evaluate the contribution of each stage, we conduct an ablation study using different combinations of the three stages, as demonstrated in Table 6. The base model, Qwen-2.5-VL-7B, demonstrates near-random performance across all datasets, indicating its limited native capability in detecting forged faces. Adding only Stage 3, which introduces user-specific tuning, moderately improves performance (e.g., +20.96 on BlendFace and +17.88 on PuLID), but results remain suboptimal due to the model’s lack of foundational face understanding. Introducing Stage 1 alongside Stage 3 yields substantial gains across all benchmarks, confirming that basic face comprehension is essential for effective user-specific customization. Combining Stages 1 and 2, without user-specific tuning, enhances general discrimination (e.g., +45.50 on BlendFace compared to the base model), but is still inferior to the full configuration. Notably, the detection for VIP users was conducted under a one-shot setting, where the MLLM compares a single reference image of the target user with the queried image. Without user-specific customization, the MLLM cannot accurately capture the nuanced identity traits of the

⁵<https://github.com/modelscope/ms-swift>

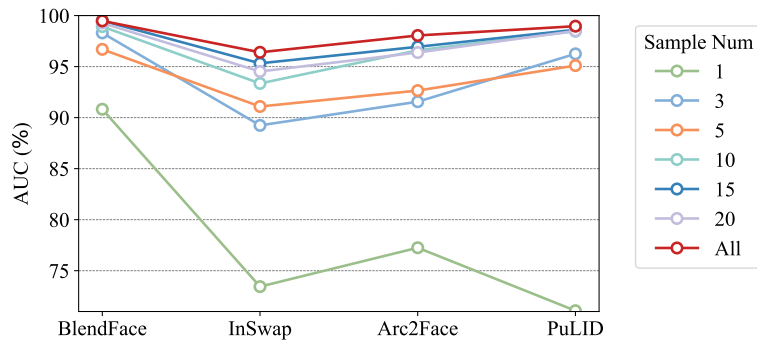


Figure 6: Ablation study on VIPGuard performance (AUC (%)) with varying numbers of available authentic images of target users. ‘All’ refers to using all available authentic images, typically ranging from 20 to 40 images for each identity in VIPBench.

VIP user, resulting in limited detection performance. The complete VIPGuard pipeline (Stages 1, 2, and 3) achieves the highest performance across all datasets, with detection rates exceeding 96% in most cases (e.g., 99.48 on BlendFace, 98.96 on PuLiD), highlighting the critical synergy among the three stages. These results demonstrate that VIPGuard’s effectiveness hinges on the sequential integration of facial understanding, general identity discrimination, and user-specific adaptation. Each component contributes uniquely, and omitting any stage leads to measurable performance degradation.

Impact of the Number of Available Authentic Images for Target Users (VIPs) While abundant negative samples can be obtained from real images of different identities, the number of accessible authentic images for a target user limits the positive samples available for Stage 3 training, thereby directly impacting the model’s user-specific detection capability. In this experiment, we varied the quantity of available authentic images for a specific user and explored its effect on VIPGuard’s performance (Figure 6). When only one authentic image was accessible, the scarcity of positive samples precluded Stage 3 training; consequently, the Stage 2 trained model was used directly for detection without User-Specific Customization. The results indicate that VIPGuard’s detection capability improves as more authentic images of the specific user become available. Compared to directly using the Stage 2 model, employing Stage 3 with even just three authentic images (positive samples) yielded a substantial improvement in detection performance. Furthermore, as the number of accessible user samples increased, detection performance across various forgery types showed consistent improvement.

Impact of the Size of VIP tokens in Stage3 In Stage 3, the VIP Token μ is of size $n \times d$, where n denotes the number of VIP Tokens and d represents the feature dimensionality. We investigated the impact of varying n on performance. As shown in Figure 7, increasing n up to 32 improved performance, reaching an optimum at $n = 32$. However, further increasing n to 64 or 128 resulted in a performance decline, potentially due to overfitting on the limited training samples. Therefore, we finally set n to 32.

Performance with Different Backbones To further validate the universality of VIPGuard, we evaluated its performance using different multimodal large language model (MLLM) backbones. As shown in Table 7, we replaced the backbone with LLaMA-3.2-Vision-11B [45], Qwen-2.5-VL-3B [1], and Qwen-2.5-VL-7B [1], respectively. Across all configurations, VIPGuard consistently achieved strong and stable performance improvements over the baseline model, demonstrating its adaptability to diverse MLLM architectures and its great detection performance across different generators.

Impact of the Face Recognition Models In the proposed method, the face recognition model is utilized to provide the global facial information and output a similarity score. To determine which face-recognition backbone is most suitable for computing similarity in VIPGuard, we conducted a comparative evaluation of several popular models, including CosFace [62], ArcFace [14], and

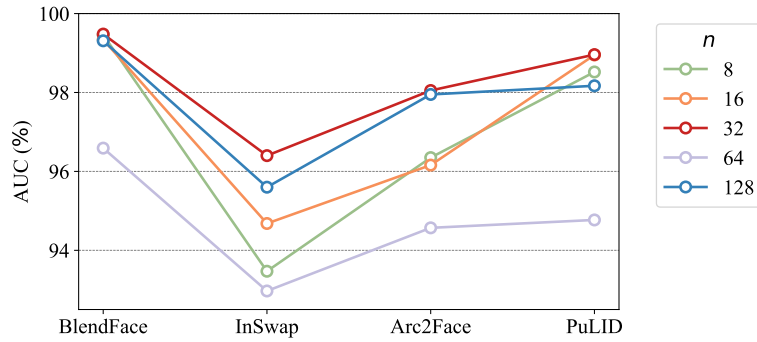


Figure 7: Ablation study on VIPGuard performance (AUC (%)) with varying the size n of the VIP token.

Table 7: Performance Comparison (AUC (%)) of VIPGuard with Different MLLMs.

Method	BlendFace	InSwap	Arc2Face	PuLID	Average
Baseline					
Qwen-2.5-VL-7B (Baseline)	51.00	49.79	49.85	50.30	50.24
VIPGuard					
LLaMA-3.2-Vision-11B	96.27	88.97	99.00	99.52	95.94
Qwen-2.5-VL-3B	90.71	87.85	87.82	97.94	91.08
Qwen-2.5-VL-7B	99.48	96.40	98.05	98.96	98.23

TransFace [11]. Each model was integrated into the VIPGuard framework to perform identity similarity estimation, and the performance was measured by AUC (%) across four representative forgery datasets: BlendFace [57], InSwap [30], Arc2Face [49], and PuLID [24]. As summarized in Table 8, TransFace consistently achieved the highest accuracy among the tested models, demonstrating its superior capability in capturing discriminative and fine-grained identity representations. Therefore, we adopt TransFace as the default face-recognition model in our experiments.

Table 8: Performance (AUC (%)) of VIPGuard using different face-recognition models as similarity components. TransFace exhibits the best overall performance and is thus used by default in our framework.

Face Model	BlendFace	InSwap	Arc2Face	PuLID	Average
CosFace	95.12	83.32	83.31	91.74	88.37
ArcFace	98.14	83.17	94.41	97.88	93.40
TransFace	99.48	96.40	98.05	98.96	98.23

D.3 Adaptive VIP Token Selection

In practical scenarios, it is common to encounter multiple VIP users requiring protection, which poses the challenge of automatically identifying and selecting the appropriate VIP token without manual effort. To address this, we develop an adaptive variant, Adaptive VIPGuard, that autonomously selects the most relevant VIP token, thereby improving the scalability and usability of the proposed framework in real-world applications. Specifically, we first leverage a face recognition model to identify the VIP user whose facial features are most similar to those in the input image. Subsequently, Adaptive VIPGuard utilizes the corresponding VIP token to perform identity-specific forgery detection. As shown in Table 9, the comparable performance between Adaptive VIPGuard and the original VIPGuard demonstrates the robustness and practicality of our framework in real-world applications.

Table 9: Comparison of AUC (%) for adaptive VIP token selection in VIPEval, where Adaptive VIPGuard refers to the automatic selection of the VIP token without manual intervention.

Method	BlendFace	InSwap	Arc2Face	PuLID	Average
VIPGuard	99.48	96.40	98.05	98.96	98.23
Adaptive VIPGuard	99.31	96.14	97.63	98.95	98.01

D.4 Robustness Evaluation

We conducted experiments under common image degradations, including Gaussian noise, Gaussian blurring, and JPEG compression, to evaluate the robustness of VIPGuard. As shown in Table 10, VIPGuard remains highly effective even under severe degradation levels. Because the detector primarily relies on high-level semantic cues such as facial structure and identity-related features, it is inherently less sensitive to low-level pixel distortions. Table 10 presents the average AUC performance of VIPGuard in detecting BlendFace [57], InSwap [30], Arc2Face [49], and PuLID [24] forgeries under different degradation intensities. The results demonstrate minimal performance decline across all settings, confirming the model’s robustness and strong generalization capacity. The detailed degradation configurations are provided in Table 11. These results confirm that VIPGuard maintains stable detection performance even when subjected to considerable image degradation, demonstrating its robustness and practicality for real-world deployment.

Table 10: VIPGuard’s performance (AUC (%)) under different image degradations. Higher levels indicate stronger degradation. The results represent the average AUC across BlendFace, InSwap, Arc2Face, and PuLID.

Level	Gaussian Noise (Color)	Gaussian Blurring	JPEG Compression
None	98.23	98.23	98.23
1	97.07	98.17	98.03
2	96.53	98.16	98.10
3	94.12	98.05	97.78

Table 11: Degradation configurations for robustness evaluation. Gaussian noise ($\mathcal{N}(0, \sigma^2)$) was applied in YCbCr space; Gaussian blurring was defined by kernel size (K) and standard deviation (σ); JPEG compression was applied using different quality factors (QF).

Level	Gaussian Noise (Color)	Gaussian Blurring	JPEG Compression
1	$\mathcal{N}(0, 8^2)$	$K = (7, 7), \sigma = 1$	QF 90
2	$\mathcal{N}(0, 11^2)$	$K = (13, 13), \sigma = 2$	QF 60
3	$\mathcal{N}(0, 18^2)$	$K = (21, 21), \sigma = 3$	QF 30

D.5 Explanation Examples

We additionally supplement some explanation results, which are exhibited in Figures 8, 9, 10, and 11.



Image 1
(Real)



Image 2
(Real)



User

Can you describe the face's structure, skin condition, and overall appearance?
.....



VIP Guard



User

Please compare the two facial images and determine whether they belong to the same person. The face similarity of these two face is **93/100**. The face tokens are shown as follows, `<|face_pad|>`. The facial attributes for the first person is: {Attr_1}\n The facial attributes for the second person is: {Attr_2} You need to make a step-by-step judgment based on different facial attributes and provide your conclusion.



VIP Guard

Okay, here is the comparison based on the provided images and attributes:

Outer face region: Both images show a consistent oval face shape with a flat jaw. The proportions regarding temporal, zygomatic, and jaw widths appear identical.

Eye region: While both have bushy eyebrows, Image 1 describes 'normal' eyes with an angulus oculi medialis of 53, whereas Image 2 describes 'thin' eyes with an angulus oculi medialis of 45. This difference in classification and angle measurement is noticeable.

Nose region: Image 1 identifies the nose as 'thin', while Image 2 identifies it as 'normal'. This represents a difference in the attribute description.

Mouth region: Both images describe the mouth as having 'smile lips'. Visually, the lip shape, fullness, and curvature appear consistent between the two images.

Skin region: Although both share light skin tone, no freckles/moles, and fine lines around the eyes, Image 1 notes the presence of nasolabial folds, which are absent in Image 2 according to the descriptions.

Summary: The comparison reveals strong similarities in the fundamental structure of the outer face, mouth, and eyebrow regions. However, discrepancies exist in the specific classifications of eye type, nose type, the presence of nasolabial folds, and the expressed emotion. These differences might stem from variations in lighting, expression, angle, or subjective interpretation during attribute annotation.

Overall Conclusion: Despite the noted differences in some attribute classifications (eyes, nose, skin details), the core facial structure appears highly consistent across both images. Considering the high face similarity score of 93/100 and the visual evidence of strong resemblance, the analysis supports the statement that both images belong to the same person. The variations likely represent minor differences in image conditions or annotation nuances rather than identity differences.

[Yes] The two images are of the same person.



VIP Guard

Figure 8: An example of VIPGuard in detecting the real faces. The image 1 and image 2 are from LAION-Face [87].



Image 1
(Real)



Image 2
(Arc2Face)



User

Can you describe the face's structure, skin condition, and overall appearance?
.....



VIP Guard

Attr_1: ...

Attr_2: ...



User

Please compare the two facial images and determine whether they belong to the same person. The face similarity of these two face is **91/100**. The face tokens are shown as follows, `<|face_pad|>`. The facial attributes for the first person is: {Attr_1}\n The facial attributes for the second person is: {Attr_2} You need to make a step-by-step judgment based on different facial attributes and provide your conclusion.

Okay, here is the comparison based on the provided images and attributes:

Outer face region: [No] Both images show individuals with a similar oval face shape and flat jawline. The proportions regarding temporal, zygomatic, and jaw widths appear consistent between the two faces.

Eye region: There are noticeable differences in the eye region. **Image 1 has normal-sized eyes** with a smaller angulus oculi medialis (53), while **Image 2 features larger eyes** with a wider angulus oculi medialis (71). Although both have bushy eyebrows, the eye size and spacing differ significantly.

Nose region: A difference exists in the nose structure. **Image 1 displays a thinner nose, whereas Image 2 has a nose described as normal, which appears slightly broader, particularly at the bridge and tip.**

Mouth region: While both are described as having smile lips, the actual expression differs (smile vs. neutral). More importantly, the lip fullness and the specific curvature of the lips appear distinct between the two images.

Skin region: Differences are noted in the skin. **Image 1 lacks a mole, whereas Image 2 has one mole present. Additionally, the skin texture and pore appearance seem slightly different upon visual inspection.**

Summary: While some general facial structures like face shape and jaw type share similarities, significant differences are observed in the eye size and spacing, nose shape, lip characteristics, and skin details.

Overall Conclusion: Despite the high face similarity score of 91/100 and shared attributes like face shape and jaw type, the distinct differences identified in the eye region, nose, mouth, skin, and overall expression confirm that the two images feature different individuals.

[No] The two images are of different people.



VIP Guard

Figure 9: An example of VIPGuard in detecting the forgery generated by Arc2Face [49]. The image 1 is from LAION-Face [87] and the image 2 was generated by Arc2Face.



Image 1
(Real)



Image 2
(GPT4o)



User

Can you describe the face's structure, skin condition, and overall appearance?
.....

Attr_1: ... Attr_2: ...



VIP Guard



User

Please compare the two facial images and determine whether they belong to the same person. The face similarity of these two face is **93/100**. The face tokens are shown as follows, `<|face_pad|>`. The facial attributes for the first person is: {Attr_1}\n The facial attributes for the second person is: {Attr_2} You need to make a step-by-step judgment based on different facial attributes and provide your conclusion.

Okay, here is the comparison based on the provided images and attributes:

Outer face region: Both images show a long face shape with a square jaw. The described proportions between temporal, zygomatic, and jaw widths are consistent in both attribute sets and appear visually similar.

Eye region: Although both have big eyes and bushy eyebrows, **there is a difference in the angulus oculi medialis measurement (63 vs 71)**. Additionally, the left eye in Image 2 is occluded, whereas both eyes are open in Image 1.

Nose region: Both images feature a thick nose, and the visual appearance of the nose shape and size is consistent between the two images.

Mouth region: Both individuals display a smile lip type. The general shape, curvature, and teeth exposure during the smile appear very similar.

Skin region: While both share natural skin tone, crow's feet, eye fine lines, nasolabial folds, and eye pouches, **Image 1 has a distinct mole on the cheek, which is absent in Image 2.**

Summary: The comparison reveals strong similarities in the outer face shape, nose, and mouth. However, noticeable differences exist in the eye region (specifically the angulus oculi medialis measurement and eye occlusion) and the skin region (presence/absence of a distinct mole).

Overall Conclusion: Despite numerous similarities and a high face similarity score (93/100), the specific differences observed in the eye and skin regions confirm that the two images depict different individuals.

[No] The two images are of different people.



VIP Guard

Figure 10: An example of VIPGuard in detecting the forgery generated by GPT-4o [48]. The image 1 is from CrossFaceID [64] and the image 2 was generated by GPT-4o.

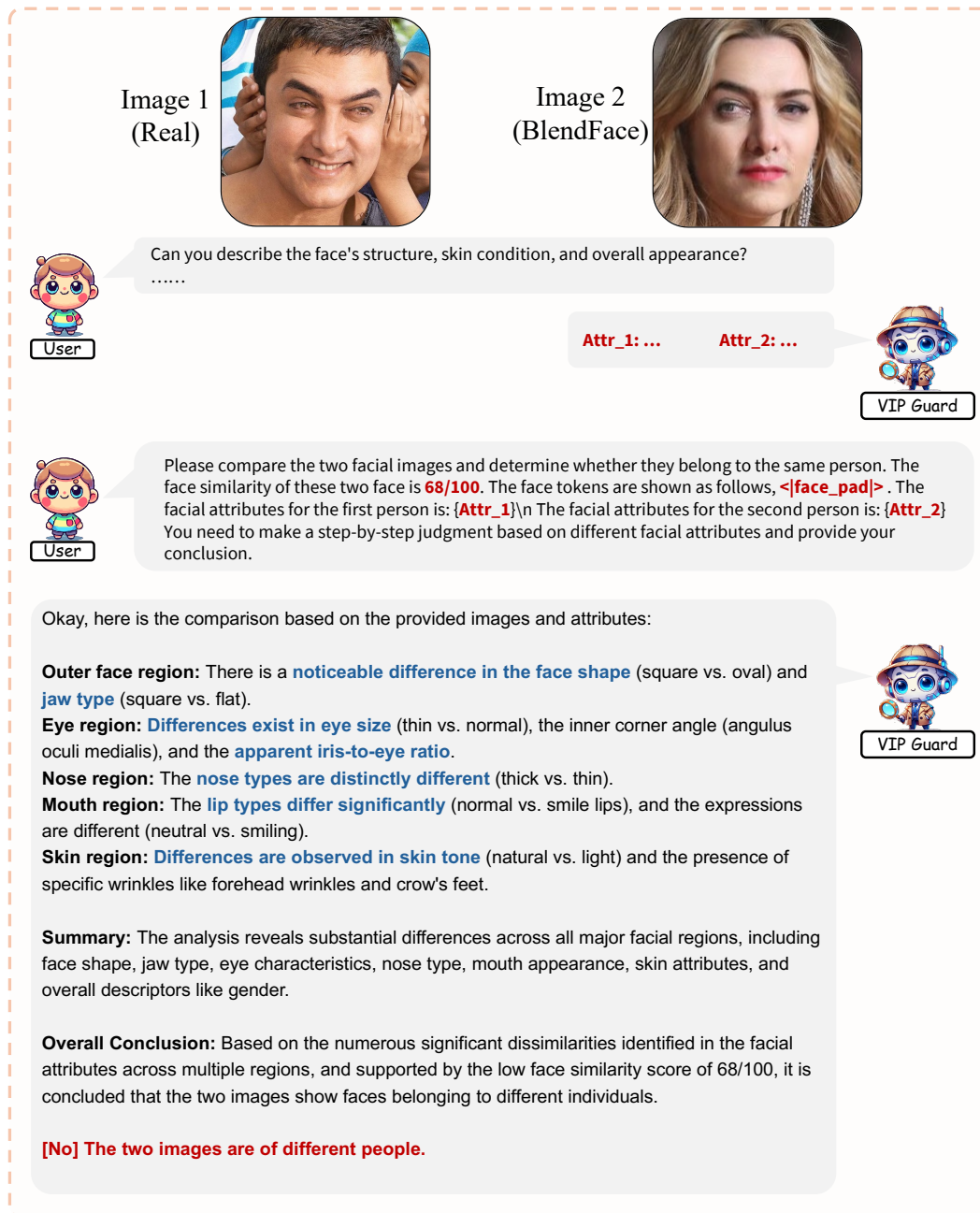


Figure 11: An example of VIPGuard in detecting the forgery generated by BlendFace [57]. The image 1 is from FaceID-6M [63] and the image 2 was generated by BlendFace.

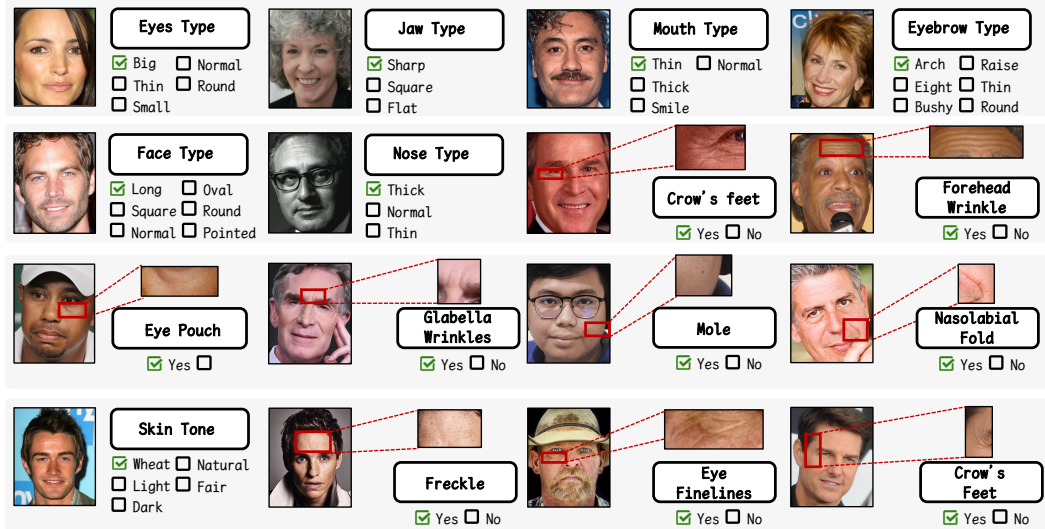


Figure 12: Visualization of different facial attributes. All the images are from LAION-Face [87]

E Dataset Construction

We introduce a new dataset, VIPBench, designed to provide a more comprehensive evaluation of Identity-aware Deepfake Detection methods [54, 19, 17, 16, 81]. In this section, we detailed describe the construction pipeline of VIPBench, which contain three part-Facial Attributes Description Dataset $\mathcal{D}_{FA}$, Identity Discrimination Dataset $\mathcal{D}_{ID}$, and VIPEval $\mathcal{D}_{Eval}$. The detailed process encompasses data collection&preprocessing and the construction of visual question answering (VQA) components, which are shown below.

E.1 Facial Attributes Description Dataset

To improve the ability of facial understanding of MLLMs, we proposed Facial Attributes Description Dataset $\mathcal{D}_{FA}$, a multimodal dataset composed of *high-resolution* facial images paired with rich *facial attribute descriptions*.

Data Collection&Preprocessing High-resolution facial images are essential for effective facial understanding, as they preserve rich and detailed information. To this end, we collected a large number of high-quality facial images (with resolution more than 1024×1024) from LAION-Face [87], filtering out those that were blurry or occluded. Specifically, we employed MTCNN [85] to crop faces from the raw images. We retained only those samples where facial landmarks could be reliably detected to ensure clarity. Finally, we selected faces with yaw and pitch angles below 15° to ensure frontal and unobstructed views by 3DDFA [23].

Construction of VQA data We adopt a commercial facial analysis tool, MegVii's official API⁶, to obtain detailed facial attributes. The facial attributes are shown in Figure 12. Subsequently, this structural information was used to build a VQA dataset that textually describes facial images. Specifically, we constructed three types of VQA data: multiple choice, short answer, and long answer. For the multiple-choice and short-answer formats, we directly constructed the VQA data using the extracted facial attributes, as illustrated in Figure 13. In addition, to generate long-answer VQA data, we employed GPT-4o [48] to reorganize the structured facial attribute information into coherent and natural language descriptions, as demonstrated in Figure 14. Figure 15 illustrates an example of the long answer format in $\mathcal{D}_{FA}$. The naive MLLM will gain improved facial understanding abilities after training on the proposed facial VQA datasets.

⁶<https://www.faceplusplus.com.cn/>

VQA: Multi-Choices

Question

Which of the following best describes the person's Skin tone?

A. Light B. Wheat C. Fair D. Natural E. Dark

Answer

A



VQA: Short Answer

Question

Describe the observed Forehead wrinkle of the person in the input image.

Answer

No.



Figure 13: Examples of our VQA data formats, including a multiple-choice question and a short-answer question based on facial attributes. The two images are from LAION-Face [87]

Prompt to GPT4o

You are a helpful assistant responsible for paraphrasing or restructuring the input text. Your goal is to improve its clarity, style, or tone without altering the original meaning. Do not fabricate or introduce any information that is not already present in the input. You must not change any attribute names or their values. For each major category, organize all its content into a coherent paragraph, and include a clear subheading for each category. You must only output the rewritten text—do not include any additional or unrelated responses.

- The attributes in skin region of face are as following:
- Skin tone: light
- Freckle: no
- Mole: yes
- Forehead wrinkle: no
- Crows feet: no
- Eye finelines: yes
- Glabella wrinkle: no
- Nasolabial fold: no
- Eye pouch: yes

Response of GPT4o

The characteristics of the facial skin region are as follows: the skin tone is light, there are no freckles, a mole is present, there are no forehead wrinkles, no crow's feet, fine lines are visible around the eyes, there are no glabella wrinkles, no nasolabial folds, and eye pouches are present.

Figure 14: An example of using GPT-4o to generate a coherent textual description by reorganizing structured facial attribute data. This case focuses on attributes related to the skin region as a representative example.

VQA: Long Answer



Question

Can you describe the face's structure, skin condition, and overall appearance?

- **Outer Face Region** (attribute list: [face type, jaw type, comparison of temporal width and zygomatic width, comparison of jaw width and zygomatic width, comparison of temporal width and jaw width])
- **Eyes Region** (attribute list: [eye type, angulus oculi medialis, eyebrow type]), and **Nose Region** (attribute list: [nose type])
- **Mouth Region** (attribute list: [mouth type]).
- **Skin Region** (attribute list: [skin tone, freckle, mole, forehead wrinkle, crows feet, eye finelines, glabella wrinkle, nasolabial fold, eye pouch]).
- **Overall facial description** (attribute list: [gender, smile, glass, head position (up/down), head position (left/right), eye visibility and glasses detection, nose visibility, chin visibility, face visibility, emotion, face quality])

Answer

Outer face region: In the outer face region, the attributes include an oval face type and a flat jaw type. the comparison of temporal width and zygomatic width shows them to be similar, while the jaw width is narrower compared to the zygomatic width. additionally, the temporal width is wider in comparison to the jaw width.

Eye region: For the eyes region, the face features big eyes, with an angulus oculi medialis measurement of 60. the eyebrows are described as bushy.

Nose region: The nose region is characterized by a thick nose type.

Mouth region: In the mouth region, the type is categorized as smile lip.

Skin region: The skin attributes include a wheat skin tone, with no freckles or moles present. the forehead is wrinkle-free, while crows feet and fine lines around the eyes are noticeable. there are no glabella wrinkles, and nasolabial folds are present, but there is no eye pouch.

Overall facial description: The overall description of the face includes the following characteristics: gender is female, exhibiting a smile, and there are no glasses worn. the head position is normal when facing forward. both eyes are visible, open, and not wearing glasses. the nose and chin are clear, and both the left and right sides of the face are visible. the emotion displayed is happiness, and the face quality is described as very clear, making it suitable for face verification.

Figure 15: An example of long-answer VQA data. The images come from LAION-Face [87]

E.2 ID Discrimination Dataset

In this work, we formulate the Deepfake detection task as a fine-grained face recognition problem tailored to specific users. To support this identity verification task, we introduce two datasets: $\mathcal{D}_{ID}^{general}$ and $\mathcal{D}_{ID}^{vip}$. The former consists of facial image pairs drawn from arbitrary identities, whereas the latter comprises facial image pairs constructed specifically around a particular user.

Data Collection&Preprocessing To construct these datasets, we first require a large volume of facial images with known identities. To this end, we source relevant images from existing open source datasets, including LAION-Face [87], FaceID-6M [63], and CrossFaceID [64]. As illustrated in Figure 16, we utilized the image-caption pairs from these open source datasets and employed Deepseek to extract names from the captions, thereby generating a pool $\mathcal{J}$ of image-name pairs. We retrieved multiple images for each identity based on the extracted names and applied the same image pre-processing techniques as described in Section E.1. Subsequently, as illustrated in Figure 17, we constructed facial image pairs for both $\mathcal{D}_{ID}^{general}$ and $\mathcal{D}_{ID}^{vip}$. The collected images were organized into a large number of pairs, each comprising a reference image I^r and a test image I^t . For $\mathcal{D}_{ID}^{general}$, positive pairs comprise two real images of the same identity (denoted as $I_{real}^r - I_{real}^t$ -Same ID), while negative pairs include either two real images from different identities ($I_{real}^r - I_{real}^t$ -Diff ID) or a real image paired with its corresponding fake counterpart ($I_{real}^r - I_{fake}^t$ -Same ID). To generate fake images, we employed SimSwap [5] and Arc2Face [49]. Notably, when using SimSwap, we replaced the identity vector typically required for face swapping with a random noise vector $\sigma \sim \mathcal{N}(0, 1)$, enabling the substitution of the inner face while preserving the outer facial features. In addition, we constructed partially swapped images by applying different masks to the eyes, nose, mouth, and inner face regions. The ratio of samples in $\mathcal{D}_{ID}^{general}$ is

$$(I_{real}^r - I_{real}^t \text{-Same ID}) : (I_{real}^r - I_{real}^t \text{-Diff ID}) : (I_{real}^r - I_{fake}^t \text{-Same ID}) = 2 : 1 : 1,$$

which ensures a balanced number of positive and negative samples.

Similar to $\mathcal{D}_{ID}^{general}$, the construction pipeline for $\mathcal{D}_{ID}^{vip}$ followed a nearly identical procedure, with the key distinction that all reference images I^r exclusively belonged to VIP users. Due to the few number of the real images of VIP users, the ratio was adjusted to

$$(I_{real}^r - I_{real}^t \text{-Same ID}) : (I_{real}^r - I_{real}^t \text{-Diff ID}) : (I_{real}^r - I_{fake}^t \text{-Same ID}) = 1 : 5 : 5,$$

where the number of negative samples is ten times that of the positive samples.

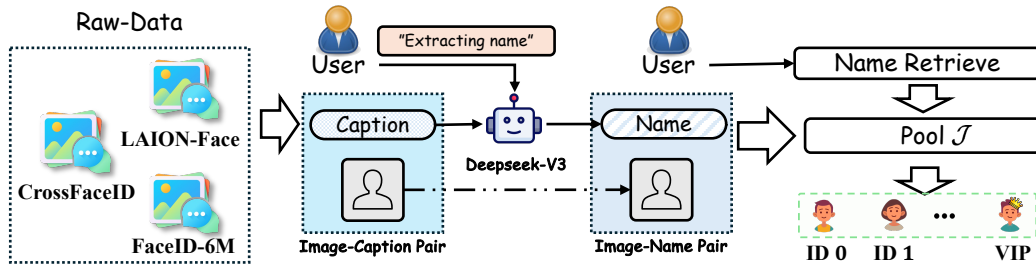


Figure 16: Pipeline for constructing the image-name pair pool $\mathcal{J}$.

Construction of VQA data After constructing the facial image pair dataset, we first annotate facial attributes using a captioning model (Qwen-2.5-VL-7B), which was fine-tuned on the $\mathcal{D}_{FA}$ dataset. These attribute annotations, together with facial similarity scores, are then provided as input to Gemini⁷ [21], which is tasked with analyzing the similarities and differences between the two faces. Figure 18 illustrates an example of the prompt used during this process. The prompts are adjusted based on the sample type. For positive samples, Gemini is explicitly instructed with the note: “Note that these two images show the same person.” For negative samples, the prompt states: “Note that these two images show different persons.” Furthermore, in cases

⁷Gemini API version in use: 2.5-pro-exp-03-25.

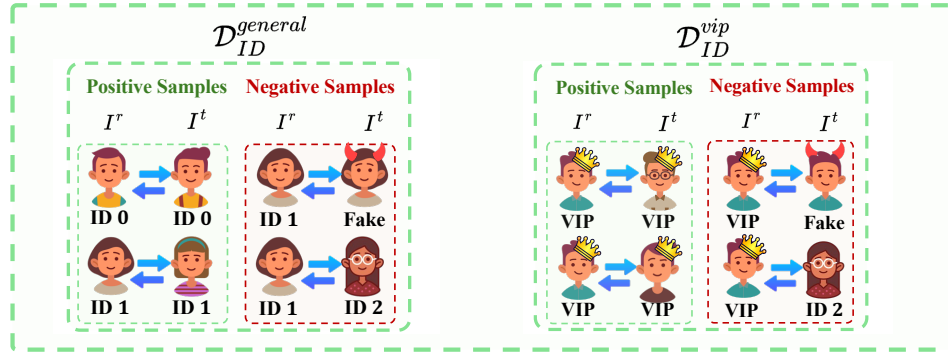


Figure 17: The composition of facial pairs in $\mathcal{D}_{ID}^{general}$ and $\mathcal{D}_{ID}^{vip}$.

where the negative sample involves a fake image, we provide a more nuanced prompt: “Although the faces may appear similar, they are not the same person.” Importantly, Gemini was constrained to base its reasoning strictly on the provided attribute annotations, thereby mitigating hallucinations. Finally, the VQA data and the corresponding facial image pairs constituted our dataset $\mathcal{D}_{ID}^{general}$ and $\mathcal{D}_{ID}^{vip}$.

E.3 VIPEval: User-Specific Evaluation Dataset

To better evaluate the effectiveness of user-specific protection methods and ID-aware detectors, we constructed a multi-resolution dataset that includes seven types of face swapping and seven types of fully synthetic face forgeries. In this section, we present representative samples from the dataset. In Figure 19, we present a comparison between authentic images from CelebDF [39] and our VIPEval dataset. VIPEval includes a large number of images with various resolutions for each identity that capture significantly more facial detail.

Data Collection&Preprocessing The data collection and pre-processing procedure in VIPEval follows the same steps as described in Section E.1. Building on this, we collected images for 22 target identities, obtaining approximately 40–60 real images per identity. For each identity, 20 real images were selected and fixed as the real samples in the test set, as well as the target images for generating fake content. The remaining 20–40 images were used as training data for that identity, and were used to construct $\mathcal{D}_{ID}^{vip}$ following the method described in Section E.2. The training set $\mathcal{D}_{ID}^{vip}$ is individually customized for each identity.

Prompt to Gemini

You will receive two input images with the same identity:

- Image 1: The face of the primary (reference) user. - Image 2: The face to be compared.
- The face similarity score between Image 1 and Image 2 is {face_score}/100.
- **Note that these two images show the same person.**
- Now, the region information of facial attribute of these two images is provided as follows: For Image 1: {attr_1} For Image 2: {attr_2}
- According to the input images, you can supplement additional facial attribute information (such as the ratio between the iris size and the whole eye size in the eye region, the curvature of the lips in mouth region, the variations in pore size and appearance, and so on) and comparisons based on your understanding in the specified facial regions above. For each region, first provide a conclusion by stating either [Yes] if there is a noticeable difference, or [No] if there is no significant difference. Then explain the specific similarities or differences based on the facial attributes for that region. If the conclusion is [No], you should give a more detailed explanation. If the conclusion is [Yes], you must provide a simpler explanation. After completing the region-by-region analysis, write a short summary and overall conclusion. You may also refer to the face similarity score as supporting evidence in your analysis.
- Note: There may be more than one region that shows dissimilarity.
- Note: You do not need to identify the specific person—your task is to analyze and compare based solely on facial attributes.
- Note: You must only output relevant responses without any unrelated information.
- Note: Respond using conversational text only—do not return JSON or code format.
- Note: You should provide a precise answer as short as possible.
- Note: You should only provide the comparison process but not list the facial attributes.

Response of Gemini

Okay, here is the comparison based on the provided images and attributes:

1. **Outer face region:** [Yes] While the jaw type and relative width comparisons (temporal/zygomatic/jaw) are consistent, the described face type differs (long vs. square). Visually, the overall shape appears very similar, suggesting the difference might be due to slight angle variations or detection nuances.
2. **Eye region:** [Yes] Although the eye type (big) and eyebrow type (bushy) match, there is a noted difference in the angulus oculi medialis measurement (63 vs. 71). Visually, the eyes, iris size relative to the eye, and eyebrows appear highly consistent in shape and placement.
3. **Nose region:** [No] The nose type is consistently identified as thick in both descriptions. Visually, the shape, bridge structure, and nostril appearance are very similar across both images.
4. **Mouth region:** [No] The mouth type is described as smile lip in both instances. The shape, curvature, and thickness of the lips appear consistent in the images.
5. **Skin region:** [No] Both descriptions indicate a natural skin tone without freckles or moles, but with fine lines around the eyes, nasolabial folds, and eye pouches. These features, along with the general skin texture, appear consistent between the images, accounting for minor lighting variations.
6. **Overall facial description:** [No] All aspects match: male gender, neutral expression, no glasses, normal head pose, clear visibility of features, and very clear face quality suitable for verification.

Summary: The comparison reveals consistency across most facial regions, including the nose, mouth, skin characteristics, and overall description. Minor discrepancies were noted in the outer face type description and the eye angle measurement, though visual inspection shows strong similarities in these areas as well.

Overall Conclusion: Despite the minor noted differences in specific attribute descriptions (face type, eye angle), the overwhelming similarity in major facial features (eyes, nose, mouth shape), skin details, and overall appearance, strongly supported by the high face similarity score of 91/100, indicates that the two images belong to the same person.

Figure 18: An instance of using Gemini to generate a textual description for facial discrimination. face_score is the face similarity score calculated by face recognition models. attr_1 and attr_2, formatted as shown in the answer in Figure 15, should be filled with attribute information generated by the captioning model.

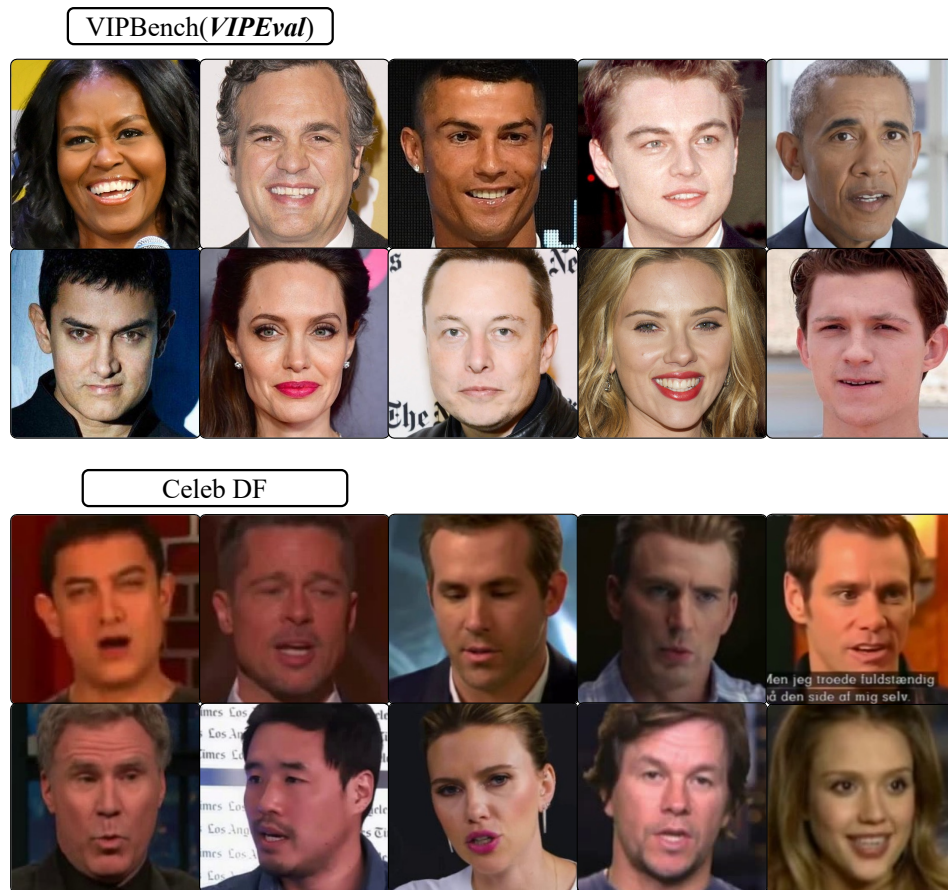


Figure 19: Comparasion of VIPBench (Ours) and CelebDF [39]. The quality of authentic images is higher in VIPBench. The images in VIPBench are from LAION-Face [87], FaceID-6M [63], and CrossFaceID [64].

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We propose a method named VIPGuard and a dataset named VIPBench. It is exactly matches the experiment in this paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss our limitations in Supplementary material ("Limitations").

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our work is purely empirical and does not include any formal theoretical results or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We will provide comprehensive experimental details in the supplemental material. All experiments in this paper are reproducible.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release the dataset and code upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We append all the experimental details and dataset setting in the supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We report metrics averaged over multiple independent runs to support robustness, but have omitted explicit error bars in this version due to space constraints; we will include full variability analysis with standard deviations in the supplemental material and public code release.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: Detailed compute specifications (hardware type, memory usage) are not included in this submission.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We confirm that our research fully adheres to the NeurIPS Code of Ethics

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The impacts will be discussed in the supplementary material

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [\[Yes\]](#)

Justification: This paper proposes a dataset involving facial images. To strictly prevent misuse, the dataset will be managed through a request form (hosted on HuggingFace) to monitor and control its usage.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [\[Yes\]](#)

Justification: We reference all the open-source models including Qwen2.5VL, LLaMA 3.2 Vision, and so on—including their paper and original sources.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: We introduce a New Deepfake dataset and provide its detailed schema, usage instructions, and metadata in Appendix; full documentation will accompany the public release.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: Our work does not involve any crowdsourcing experiments or human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: Our work does not involve any human subjects or user studies.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [\[Yes\]](#)

Justification: We used Qwen2.5VL as our foundation model.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.