# Identity-Aware Vision-Language Model for Explainable Face Forgery Detection

**Junhao Xu**[1, 3], **Jingjing Chen**[1, 2*], **Yang Jiao**[1], **Jiacheng Zhang**[1], **Zhiyu Tan**[1, 3]
**Hao Li**[1, 3], **Yu-Gang Jiang**[2],

[1] College of Computer Science and Artificial Intelligence, Fudan University
[2]Institute of Trustworthy Embodied AI, Fudan University
[3]Shanghai Academy of Artificial Intelligence for Science
junhaoxu23@m.fudan.edu.cn, chenjingjing@fudan.edu.cn, yjiao23@m.fudan.edu.cn,
jiachengzhang22@m.fudan.edu.cn, tanzhiyu@sais.org.cn, lihao_lh@fudan.edu.cn, ygj@fudan.edu.cn

## Abstract

Recent advances in generative artificial intelligence have enabled the creation of highly realistic image forgeries, raising significant concerns about digital media authenticity. While existing detection methods demonstrate promising results on benchmark datasets, they face critical limitations in real-world applications. First, existing detectors typically fail to detect semantic inconsistencies with the person's identity, such as implausible behaviors or incompatible environmental contexts in given images. Second, these methods rely heavily on low-level visual cues, making them effective for known forgeries but less reliable against new or unseen manipulation techniques. To address these challenges, we present a novel personalized vision-language model (VLM) that integrates low-level visual artifact analysis and high-level semantic inconsistency detection. Unlike previous VLM-based methods, our approach avoids resource-intensive supervised fine-tuning that often struggles to preserve distinct identity characteristics. Instead, we employ a lightweight method that dynamically encodes identity-specific information into specialized identifier tokens. This design enables the model to learn distinct identity characteristics while maintaining robust generalization capabilities. We further enhance detection capabilities through a lightweight detection adapter that extracts fine-grained information from shallow features of the vision encoder, preserving critical low-level evidence. Comprehensive experiments demonstrate that our approach achieves 94.25% accuracy and 94.08% F1 score, outperforming both traditional forgery detectors and general VLMs while requiring only 10 extra tokens.

## Introduction

Recent advances in generative models, particularly GAN-based deepfakes and conditioned diffusion models, have enhanced image manipulation capabilities. They enable the effortless modification of identities, actions, and contextual elements with unprecedented realism and accessibility. These advancements raise significant concerns about media authenticity and the potential spread of visual misinformation.

In response to these emerging threats, the research community has developed various face forgery detection approaches, typically leveraging CNN-based and Vision Transformer architectures to identify manipulation artifacts. Despite promising results in benchmark datasets, current detection methods face three critical limitations that hinder their effectiveness in real-world applications. First, existing approaches are unlikely to detect semantic inconsistencies related to identity, such as implausible behaviors, incompatible clothing, or incongruous environmental contexts for a specific person. As illustrated in Fig. 1, while conventional detectors may identify known manipulation artifacts (Case 1), they struggle with novel forgery methods (Case 2) where the manipulation appears visually plausible but contains semantic inconsistencies with the person's identity. Second, existing detectors frequently struggle with generalization ability, exhibiting significant performance degradation when confronted with novel forgery techniques not represented in training data. These limitations arise from an overreliance on low-level visual artifacts specific to certain generation al-

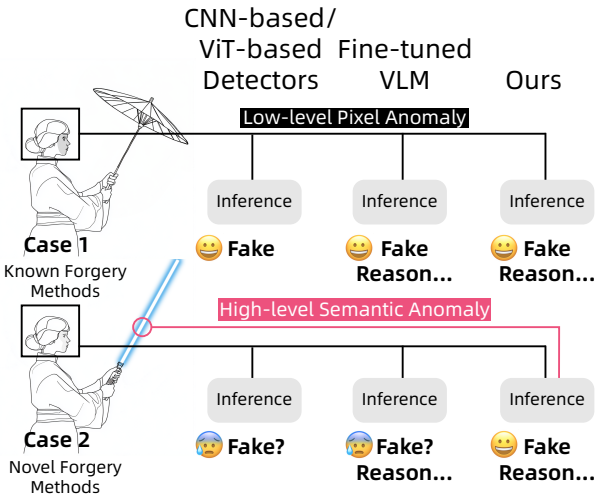


Figure 1: While existing detectors and fine-tuned VLMs can handle known manipulation methods (Case 1), they usually struggle with novel forgery techniques (Case 2). Our approach leverages both low-level artifact detection and high-level semantic analysis through personalized identity priors, enabling robust detection with explanatory reasoning for both scenarios.

*Corresponding author.

gorithms, while overlooking higher-level semantic inconsistencies that could provide more robust detection cues.

Given these limitations, vision-language models (VLMs) offer promising capabilities for addressing the challenges in face forgery detection through their inherent high-level semantic modeling and cross-modal reasoning abilities. However, naively applying supervised fine-tuning to VLMs for forgery detection presents significant challenges. First, while supervised fine-tuning offers a straightforward approach to enhance model performance in detecting image forgery, the simultaneous modeling of multiple individuals increases the challenge of preserving distinct identity characteristics, which leads to excessively additional annotated data for effective memorization and modeling of identity-specific representations. Second, real-world deployment frequently necessitates continuous integration of new individuals, requiring persistent memorization of emerging identities. The conventional paradigm of recurrent model upgrade through direct fine-tuning not only imposes prohibitive temporal and computational burdens, but also introduces the risk of potential forgetting of previously learned representations.

To address these limitations, we introduce a novel personalized vision-language model built upon the LLaVA(Liu et al. 2023) framework to integrate both low-level artifact analysis and high-level personalized anomaly detection for comprehensive image forensics detection. Our approach requires only a small set of authentic reference images of an individual to embed forgery-sensitive prior knowledge into special identifier tokens and soft tokens. We implement a lightweight strategy through personalized identifier tokens that effectively capture individual-specific visual characteristics, including appearance and behavior aspects. In contrast to supervised fine-tuning approaches, our method dynamically encodes identity-specific information. This strategy enables the model to learn distinct identity characteristics while maintaining generalization capabilities across diverse forgery detection scenarios.

The key insight driving our approach is that by combining injected personalized identity priors with VLMs' inherent commonsense reasoning capabilities, the model can effectively leverage high-level semantic cues for forgery detection. Our special identifiers serve dual roles: (1) enabling the VLM to identify identity-inconsistent information from input images and (2) guiding the generation of detailed, evidence-based forgery analysis explanations. To enhance the detection of low-level manipulation artifacts, we design a lightweight detection adapter that extracts fine-grained information from the shallow features of the VLM's vision encoder. This architectural enhancement enables more accurate and robust forgery detection across diverse manipulation techniques. It preserves critical low-level evidence typically dropped in deep feature hierarchies optimized for semantic understanding.

Our main contributions include:

- **Personalized Identity Priors**: We introduce an effective method for encoding identity priors into minimal learnable parameters to personalize VLMs. Our method enables recognition of identity-inconsistent elements without extensive fine-tuning or large-scale annotation.
- **Multi-level Feature Integration**: We propose a lightweight Detection Adapter that enhances the VLM's ability to capture low-level visual artifacts often overlooked by standard visual encoders while preserving the model's capacity to identify high-level semantic inconsistencies.
- **Comprehensive Evaluation**: Extensive experiments demonstrate that our approach achieves 94.25% accuracy and 94.08% F1 score, outperforming both traditional forgery detectors and general VLMs while requiring only 10 extra tokens.

## Related Work

### Face Forgery Detection

Most face forgery detection methods focus on detecting visual defects within facial regions. Beyond defect-based detection, researchers have explored identity-related cues for robust detection. Dong *et al.* (Dong et al. 2020) used identity-associated attributes to learn discriminative embeddings, while Cozzolino *et al.* (Cozzolino et al. 2021) developed an identity-aware framework leveraging 3D morphable models for threshold-based discrimination. Dong *et al.* (Dong et al. 2022) proposed an identity consistency transformer modeling representations across facial regions. Huang *et al.* (Huang et al. 2023) captured explicit and implicit identity information through contrastive and exploration losses. Xu *et al.* (Xu et al. 2024) proposed a reference-assisted framework using dual contrastive learning for identity-sensitive features and cross-modal inconsistencies. Despite promising results, these methods function as black-box systems, lacking transparent explanatory capabilities necessary for forensic applications. This highlights the need for approaches providing both detection and interpretable explanations.

### Vision-Language Models for Detection

VLMs integrate visual perception with language understanding, enabling cross-modal reasoning and explanation generation. However, VLMs typically lack specialized mechanisms for face forgery detection. Jia *et al.* (Jia et al. 2024) found that GPT-4V demonstrated only moderate detection capability with notable challenges on real images, performing worse than existing detection methods. Niki *et al.* (Foteinopoulou, Ghorbel, and Aouada 2024) evaluated four VLMs (BLIP-2, InstructBLIP, InstructBLIP-XXL, LLaVa-1.5) and revealed considerable limitations in generalization.

To address these limitations, specialized VLM-based approaches have emerged. Zhang *et al.* (Zhang et al. 2024) used crowdsourced human annotations to fine-tune multimodal models like BLIP. Recent frameworks including FFAA (Huang et al. 2024) and FakeShield (Xu et al. 2025) leveraged GPT-4o to generate large-scale annotations for model optimization.

Current VLM-based approaches face two fundamental limitations. First, they focus on detecting low-level visual artifacts through supervised fine-tuning, deviating from

VLMs' inherent strengths in semantic reasoning. Second, they rely on base VLMs that struggle with visual artifacts to generate training data, introducing false supervision signals treated as ground truth.

Our approach addresses these issues by leveraging VLMs' natural strength in semantic understanding to detect identity inconsistencies rather than relying on artifact identification. Inspired by VLM personalization research (Nguyen et al. 2024; Alaluf et al. 2024), we encourage our VLM to use identity-related priors for detection. This strategy aligns with identity-aware methods while offering two advantages: (1) preserving original reasoning capabilities, and (2) enabling interpretable explanations for detected forgeries.

# Methodology

## Problem Formulation

Given an input image $I$ of a specific identity, the goal is to determine whether it is authentic or manipulated and generate a textual explanation $T$ to justify the decision. We define our task as learning a function $f$ that maps:

$$f : (I, \{R_1, R_2, ..., R_n\}) \rightarrow (y, T), \tag{1}$$

where $y \in \{0, 1\}$ indicates whether the image is real (0) or fake (1), $\{R_1, R_2, ..., R_n\}$ represents a small set of authentic reference images of the target identity, and $T$ is a detailed textual explanation outlining the evidence supporting the decision.

## Overview of Our Approach

We propose a personalized vision-language model that integrates both low-level artifact analysis and high-level semantic inconsistency detection for comprehensive forgery detection. As illustrated in Fig. 2, our framework consists of two key points: (1) Personalized Identity Prior Injection: We encode identity-specific knowledge into specialized identifier tokens that enable the model to identify inconsistencies in appearance and behavior. (2) Detection Adapter: A lightweight module that extracts fine-grained information from shallow layers of the vision encoder to preserve critical low-level evidence.

## Personalized Identity Prior Injection

Unlike previous VLM-based forgery detection methods that operate on generic visual patterns, our approach embeds personalized identity priors that enable the model to recognize identity-specific inconsistencies. Our method aligns with the philosophy of research of DreamBooth (Ruiz et al. 2023) , YoLLaVA (Nguyen et al. 2024), and Textual Inversion (Gal et al. 2023), where personalized information is encoded into minimal learnable parameters. However, unlike these generation-focused methods, we redesign this concept for the distinct challenges of the discriminative task of forgery detection by decoupling identity priors into 'appearance' and 'behavior'.

**Representation of Identity Priors.** We represent identity priors using two specialized identifier tokens, denoted as <id_a> and <id_b>, which encode appearance prior and behavioral prior, respectively:

- <id_a>: Encodes appearance-related attributes of the target identity, including facial features, hairstyle, clothing preferences, and other visual characteristics.
- <id_b>: Captures behavioral patterns, typical contexts, and plausible activities associated with the target identity.

As shown in Fig.2, each identifier is assigned $N$ learnable soft tokens that encode a distributed representation of the identity's visual characteristics:

$$\begin{aligned} &\texttt{<id\_a>} \rightarrow \{\texttt{<token\_a}_1\texttt{>}, \texttt{<token\_a}_2\texttt{>}, ..., \texttt{<token\_a}_N\texttt{>}\} \\ &\texttt{<id\_b>} \rightarrow \{\texttt{<token\_b}_1\texttt{>}, \texttt{<token\_b}_2\texttt{>}, ..., \texttt{<token\_b}_N\texttt{>}\}. \end{aligned} \tag{2}$$

This representation of identity priors offers several advantages over conventional text descriptions. First, soft tokens can capture subtle visual details that are difficult to express in language. Second, their distributed nature enables more precise discrimination between authentic and manipulated images of the target identity.

**Learning Identity Priors**. To embed identity-specific knowledge into these tokens, we utilize a small set of authentic reference images $\{R_1, R_2, ..., R_n\}$ of the target identity. The training process involves two types of tasks: (1) Appearance Recognition: Given an image, the model learns to determine whether it contains the target appearance by referencing the learned token representations. (2) Behavior Recognition: The model learns to answer questions about behaviors of the target individual, encoding behavior-related information into the specialized tokens.

Formally, our trainable parameters include the specialized identifier tokens, their associated soft tokens, and corresponding weights in the LM head:

$$\theta_{\text{prior}} = \{\texttt{<id\_a>}, \texttt{<id\_b>}, \{\texttt{<token\_a/b}_i\texttt{>}\}_{i=1}^{N}, W_{\text{new}}\}, \tag{3}$$

where $W_{\text{new}}$ denotes the newly added rows in the output embedding matrix for the extended vocabulary tokens.

## Detection Adapter for Low-Level Artifact Preservation

While specialized identifier tokens enable high-level semantic inconsistency detection, effectively identifying manipulation artifacts requires preserving low-level visual features that are often dropped in deeper layers of vision encoders. Rather than fine-tuning the entire vision encoder or introducing stronger vision models, we propose a lightweight Detection Adapter that extracts and preserves these critical low-level signals.

**Architecture Design**. Our Detection Adapter extracts features from a shallow layer of the vision encoder and projects them into visual tokens that can be directly processed by the language model:

$$T_{\text{adapter}} = \text{Proj}(F_{\text{shallow}}), \tag{4}$$

where $F_{\text{shallow}}$ represents features from the initial layers of the vision encoder, and Proj maps the features to token embeddings compatible with the language model. Such a design offers several advantages: (1) Preservation of Low-Level Evidence: shallow layers of vision transformers retain critical information about compression artifacts, noise

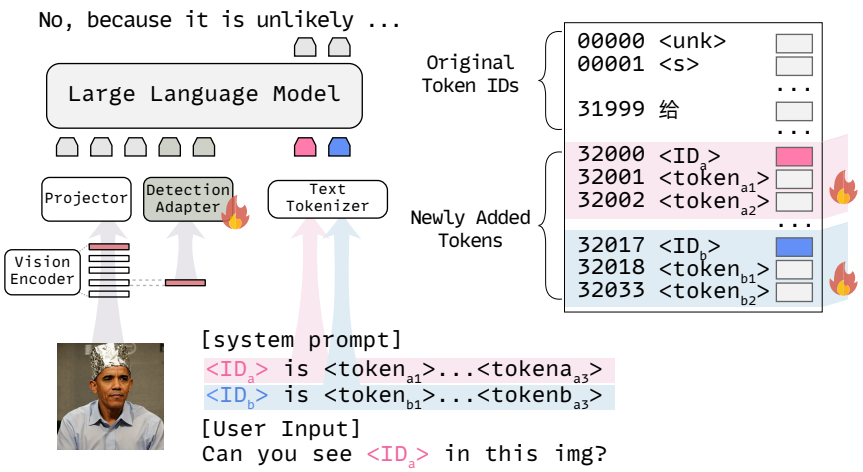


Figure 2: Overview of our proposed framework. Given a query image and a small set of authentic reference images, our approach (1) extracts personalized identity priors encoding both appearance and behavioral characteristics through specialized tokens (<id_a> and <id_b>), (2) leverages a lightweight Detection Adapter that preserves low-level visual artifacts from shallow layers of the vision encoder. The model can identify both visual inconsistencies (e.g., unnatural blending boundaries) and semantic implausibilities (e.g., inconsistent clothing or behavior) through this multi-level feature integration.

patterns, and local inconsistencies that are typically strong indicators of manipulation. (2) Parameter Efficiency: the adapter introduces minimal additional parameters compared to fine-tuning the entire vision encoder or introducing a separate model. (3) Preservation of Semantic Understanding: by maintaining the original parameters of the VLM, we preserve its strong semantic understanding capabilities while enhancing its sensitivity to forgery artifacts.

The adapter tokens $T_{\text{adapter}}$ are combined with the standard visual tokens $T_{\text{standard}}$ (derived from the original vision-language projection) and fed into the language model:

$$T_{\text{integrated}} = [T_{\text{standard}}; T_{\text{adapter}}], \tag{5}$$

where $[;]$ denotes concatenation along the sequence dimension. This allows the language model to leverage both high-level semantic information and low-level visual artifacts when generating forgery determinations and explanations.

## Training Strategy

Training our model involves two primary stages: (1) learning the detection adapter and (2) personalizing the identity priors.

**STAGE 1: Detection Adapter Training**. To train the detection adapter, we collect a diverse dataset of authentic and manipulated images without requiring identity-specific information. This allows the model to learn generic forgery detection capabilities before personalization. We formulate this as a binary classification task with a simple Yes/No answer format to reduce training complexity:

$$\mathcal{L}_{adapter} = \text{CrossEntropy}(\text{VLM}(T_{\text{standard}}; T_{\text{adapter}}), y). \tag{6}$$

Where $T_{\text{adapter}}$ are the visual tokens produced by our adapter from shallow features, $T_{\text{standard}}$ are the standard visual tokens from the vision encoder, VLM produces a Yes/No response based on these visual tokens, and $y$ is the binary ground truth label. This binary response format simplifies the optimization landscape, stabilizing training while enabling efficient reuse of existing image forgery datasets. This stage leverages readily available data without requiring identity-specific information or additional annotation efforts, as binary labels can be obtained from existing metadata.

**STAGE 2: Personalized Prior Training**. For training personalized identity priors, we create a challenging dataset with authentic images of the target identity and high-quality forgery images. These forgeries are generated using state-of-the-art diffusion-based synthesis and GAN-based face-swapping models, producing realistic manipulations to help the model develop robust discrimination capabilities. We design two types of training objectives:

- **Appearance Recognition:** a binary classification task determines if an image matches <id_a> (appearance).
- **Behavior Recognition:** a binary classification task determines if an image is consistent with <id_b> (behavior).

The training objective of this stage is:

$$\mathcal{L}_{\text{personalization}} = \mathcal{L}_{\text{appearance}} + \mathcal{L}_{\text{behavior}}. \tag{7}$$

## Data Curation

**Dataset Construction.** Due to the absence of identity-centric datasets specifically designed for VLM-based face forgery detection, we construct a novel dataset - *IDImage* that incorporates both authentic and manipulated images of target individuals, along with corresponding authenticity labels and detailed descriptions for comprehensive evaluation.

IDImage dataset comprises images of 20 diverse individuals with varying ages and genders, selected from politics and entertainment domains. These English-speaking public figures are chosen due to their substantial media presence,

making them statistically more vulnerable to digital impersonation attacks. For each individual, we curate approximately 329 real images, capturing them in diverse contexts, attire, and social settings, sometimes featuring multiple subjects within the same image to enhance complexity.

**Face Forgery Creations.** To generate face forgery creations, we employ five state-of-the-art manipulation techniques, including both traditional deepfake methods and recent conditioned diffusion-based models. Using collected real images as references, we generate face forgery creations that preserve the identity characteristics of the person while being entirely synthetic.

For the training split, we apply two widely used methods: (1) SimSwap(Chen et al. 2020), and (2) PhotoMaker(Li et al. 2024). To better evaluate generalization capability, our test split exclusively contains forgeries created using entirely different methods: (3) Roop(s0md3v 2023), (4) StoryMaker(Zhou et al. 2024), and (5) PuLID(Guo et al. 2024). The dataset contains about 279 authentic and 780 manipulated images per individual in the training split, while the test split includes about 50 authentic and 50 manipulated images per individual.

SimSwap and Roop are GAN-based face-swapping methods that transform a source face (a single reference image) into a target face while preserving the target's expressions and orientation. PhotoMaker, StoryMaker, and PuLID are conditioned diffusion-based methods that generate images guided by text prompts, using a set of reference images to maintain identity while creating new scenarios described in the prompts. The methods provided in the training set and test set are different. The deliberate distribution shift in IDImage mimics a challenging but realistic scenario, as detectors often perform poorly when training and testing on different data distributions.

# Experiments

## Experimental Setup

**Implementation Details.** In our implementation, we represent each individual with a total of 10 tokens, including identifiers <id_a>, and <id_b>, with 4 soft tokens associated with each identifier. This optimal configuration is determined through experiments. We train the tokens for each individual separately. We employ LLaVA-1.6-13B(Liu et al. 2023) as our base model, optimized using AdamW with a learning rate of 0.001. Each conversation is single-turn. The training dataset comprises approximately 279 real images per individual, along with 780 forgery images per individual (consisting of an average of 540 diffusion-generated images and 240 deepfake images). Although we use a substantial data volume per individual for a comprehensive evaluation, we later demonstrate that our approach remains effective with significantly reduced data quantities. We train each subject for only 1 epoch in about 20 minutes. All experiments are conducted on 2 NVIDIA A100 GPUs.

**Evaluation Metrics.** For forgery detectors that produce binary results, we designate real images as positive samples and both AIGC and deepfake images as face forgery creations. For all VLMs that produce natural language outputs, we use keyword retrieval to convert their responses into binary labels. Our evaluation metrics include image-level accuracy, precision, recall, and F1 score to provide a comprehensive assessment of classification performance.

## Baselines

**Forgery Detectors.** We compare our method with several representative forgery detection methods, including deepfake detectors (UCF(Yan et al. 2023), Recce(Cao et al. 2022), Xception(Chollet 2017), and Meso4(Afchar et al. 2018)), RAIRNet(Fang et al. 2023), R-MFDN(Xu et al. 2024), AI-generated content detectors (NPR(Tan et al. 2024) and ESSP(Chen, Yao, and Niu 2024)). Xception and Meso are widely used deepfake detectors, while UCF and Recce are deepfake detection methods specially designed to enhance generalization performance. RAIRNet and R-MFDN are recent deepfake detectors assisted by identity information. NPR and ESSP are recent methods that address not only GAN-generated deepfake images but also content generated by diffusion models.

**Vision-Language Models.** We also evaluate VLMs, including GPT4o-2024-08-06(Hurst et al. 2024), Qwen2.5-VL(Bai et al. 2025), and LLaVA-1.6-13B(Liu et al. 2023). Following the settings in (Jia et al. 2024), we directly query these models with the prompt: "Tell me if there are synthesis artifacts in the face or not. Must return with yes or no only." Furthermore, we compare with FFAA(Huang et al. 2024), a VLM specifically fine-tuned for face forgery detection, and Yo'LLaVA(Nguyen et al. 2024), a personalized VLM based on LLaVA. For FFAA and Yo'LLaVA, we employ the original prompts in their papers.

## Comparison with Baselines

First, we compare the performance of our method with baselines on the collected test set. To ensure a fair comparison, we fine-tune the baseline forgery detectors using the collected training data. In addition, for VLM models, we also perform supervised fine-tuning on LLaVA using the entire training split. Note that in our test set, the forgery images are generated using methods different from those seen in the training data, presenting a challenging yet more realistic and practical setting.

Table 1 shows the results. Among traditional methods, ESSP achieves the best performance (89.08% accuracy) while Meso performs the worst (55.17%). ESSP's robust noise-level analysis effectively handles unseen data, whereas Meso4 struggles with fake patterns outside its training data.

Among general VLMs, GPT-4o performs best (83.03% accuracy). LLaVA and Qwen2.5-VL perform poorly, with Qwen2.5-VL showing perfect precision but extremely low recall (12.24%), indicating accurate but infrequent fake image detection. FFAA, despite being designed for face forgery detection, achieves only 41.67% accuracy.

Fine-tuned LLaVA achieves high recall (98.08%) but moderate precision (60.71%), suggesting oversensitivity. Limited training samples cause LLaVA to recognize response patterns rather than generalizable forgery features. Yo'LLaVA shows high precision but moderate recall

| Method | ACC | Precision | Recall | F1 |
|---|---|---|---|---|
| Meso4 | 55.17 | 53.38 | **99.16** | 69.40 |
| Xception | 80.03 | 89.33 | 69.31 | 78.06 |
| Recce | 75.09 | 90.98 | 57.06 | 70.14 |
| R-MFDN† | 80.06 | 87.84 | 69.15 | 77.38 |
| UCF | 81.69 | 94.23 | 68.48 | 79.32 |
| RAIRNet† | 82.84 | 83.32 | 81.55 | 82.42 |
| NPR | 83.38 | 92.96 | 71.50 | 80.83 |
| ESSP | 89.08 | 93.18 | 83.86 | 88.27 |
| Qwen2.5-VL | 57.14 | **100.0** | 12.24 | 21.81 |
| LLaVA | 51.04 | 50.68 | 26.87 | 35.12 |
| FFAA | 41.67 | 41.58 | 47.64 | 44.36 |
| Yo'LLaVA | 75.73 | **100.0** | 50.98 | 67.53 |
| LLaVA (SFT) | 67.31 | 60.71 | 98.08 | 75.00 |
| GPT4o | 83.03 | 93.72 | 70.07 | 80.19 |
| Ours† | **94.25** | 94.68 | 94.12 | **94.08** |

Table 1: Performance comparison with baselines (%). Methods marked with† are identity-assisted detectors.

(50.98%), as it was designed for general visual recognition rather than forgery detection.

While traditional methods often outperform VLM-based models due to their domain-specific designs, VLMs offer superior usability through explainable reasoning. Our personalized approach integrates the interpretability of VLMs with enhanced detection capabilities, achieving a 94.08% F1 score and effectively identifying subtle inconsistencies in previously unseen forgery techniques.

## Comparison with Personalized VLMs

We then compare our method with VLMs by injecting personalization knowledge. To mimic a real-world scenario where a user describes a personalized subject to VLMs, we employ three methods to inject personalization knowledge for Qwen2.5-VL and GPT-4o: (1) Human-written description: human annotators manually write a description for each individual, simulating a real scenario where a user describes a personalized subject to VLMs. Each description is about 90 words ($\sim$ 150 tokens), including a rough description of appearance, clothing and styling, behavioral characteristics, and environmental context. (2) Reference Image: for each individual, we maintain a set of around 20 images featuring front-view photos without other people, clearly showing the identity. For each conversation, we randomly choose an image as a reference. (3) Reference Image + Human-written description: both the human-written description and the reference image are included in this situation. For LLaVA, we use only the human-written description to inject personalization knowledge, as it supports only a single image input.

Results in Table 2 show that personalization knowledge improves recall across all VLMs by increasing sensitivity to identity inconsistencies. However, this increases false positives, reducing precision while generally improving F1 scores. For LLaVA, human descriptions increase F1 from 35.12% to 66.69%. For Qwen2.5-VL, both descriptions

| Method | ACC | Precision | Recall | F1 | Extra Tokens |
|---|---|---|---|---|---|
| LLaVA | 51.04 | 50.68 | 26.87 | 35.12 | 0 |
| LLaVA (Text) | 51.15 | 50.08 | 99.80 | 66.69 | $\sim$ 150 |
| Qwen2.5-VL | 7.14 | **100.0** | 12.24 | 21.81 | 0 |
| Qwen2.5-VL (Text) | 56.51 | 52.89 | **100.0** | 69.18 | $\sim$ 150 |
| Qwen2.5-VL (Img) | 48.89 | 48.89 | **100.0** | 65.67 | $\sim$ 10K |
| Qwen2.5-VL (All) | 50.53 | 49.69 | **100.0** | 66.39 | $\sim$ 10K |
| GPT-4o | 83.03 | 93.72 | 70.07 | 80.19 | 0 |
| GPT-4o (Text) | 86.80 | 77.22 | 98.99 | 86.76 | $\sim$ 150 |
| GPT-4o (Img) | 65.77 | 58.88 | 99.90 | 74.09 | $\sim$ 10K |
| GPT-4o (All) | 88.02 | 80.95 | 98.98 | 89.06 | $\sim$ 10K |
| Ours | **94.25** | 94.68 | 94.12 | **94.08** | 10 |

Table 2: Comparison with Personalized VLMs (%).

and reference images boost F1 by over 44%. Text descriptions consistently outperform reference images: Qwen2.5-VL achieves 56.15% accuracy with text versus 48.89% with images, while GPT-4o achieves 86.80% with text but lower performance with images. This suggests single reference images provide limited personalization cues due to constrained viewpoints and contexts, causing overgeneralization. Combining both modalities yields mixed results: Qwen2.5-VL shows slight degradation (F1: 66.39%), suggesting difficulty integrating multimodal information, while GPT-4o achieves the best performance (F1: 89.06%), demonstrating superior multimodal alignment. Our method achieves the highest F1 score of 94.08% using only 10 extra tokens, demonstrating effective identity information utilization.

## Ablation Study

**Effect of Detection Adapter.** To evaluate the contribution of our Detection Adapter module, we compare our full model against a variant where the Detection Adapter is removed. In this ablation, input images are directly processed by the ViT encoder, producing hidden states that are transformed into visual tokens by only the original adapter. The Personalized Identity Prior Injection remains operational in both configurations. As shown in Table 3, removing the Detection Adapter (w/o Detection Adapter) increases recall but decreases other metrics, particularly precision. This suggests the model loses important low-level visual cues and becomes biased toward classifying more inputs as forgeries, reducing its overall discriminative ability.

**Effect of Decoupled Specialized Identifier Tokens.** As shown in Table 3, we empirically found that combining both appearance and behavior priors outperforms using either component alone, as they provide complementary detection cues. We also investigate whether using separate tokens for appearance and behavior priors (<id_a> and <id_b>) provides advantages over a single unified identity token (<id>). As shown in Table 3, our results demonstrate that using decoupled tokens yields better performance compared to the unified approach ((w/o Decoupled Tokens)).

**Effect of Chain-of-thought Reasoning.** We evaluated a variant that produces only binary decisions without chain-of-thought reasoning. As shown in Table 3, removing the

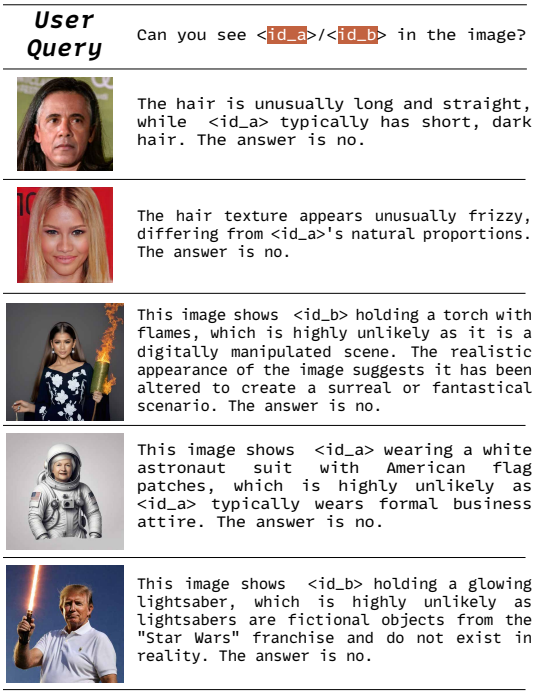


Figure 3: Testing results on IDImage dataset. More comparative results are included in the supplementary material.

| Method | ACC | Precision | Recall | F1 |
|---|---|---|---|---|
| Ours(w/o Adapter) | 82.42 | 76.56 | **98.00** | 85.96 |
| Ours(only <id_a>) | 86.03 | 79.36 | 96.88 | 87.24 |
| Ours(w/o Decoupled Tokens) | 87.09 | 88.27 | 87.62 | 86.33 |
| Ours(only <id_b>) | 87.12 | 82.58 | 93.65 | 87.77 |
| Ours(w/o CoT) | 89.09 | 93.51 | 84.03 | 87.26 |
| Ours | **94.25** | **94.68** | 94.12 | **94.08** |

Table 3: Ablation Study (%).

description (w/o CoT) in training performs worse, indicating that training the model with chain-of-thought helps learn more effective features. The reasoning process may serve as a condition for the binary decision, potentially helping the model consider relevant evidence before classification.

## Discussion on the Training Data Size

In our implementation, we use approximately 1,000 training samples per individual (279 positive samples and 780 face forgery creations) to fully explore our VLM's capabilities. However, in practical applications, collecting such extensive data for each individual is costly and labor-intensive, particularly gathering 200-300 real images per person. Therefore, we conduct experiments to evaluate our approach's performance across different data scales.

Table 4 presents the performance of our model across different training data proportions. The results show that performance improves consistently as the amount of training data increases. With only 5% of the data, the model achieves an F1 score of 68.00%, which improves substantially to 81.73% when using 10% of the data. As we continue to increase the data proportion, performance continues to improve, with the F1 score reaching 86.69% at 50% data, 90.66% at 75%, and 94.08% with the complete dataset.

| Training Data | ACC | Precision | Recall | F1 |
|---|---|---|---|---|
| 5% | 69.21 | 72.34 | 72.54 | 68.00 |
| 10% | 80.43 | 80.97 | 85.41 | 81.73 |
| 25% | 85.45 | 83.00 | 92.85 | 86.94 |
| 50% | 86.80 | 86.95 | 89.17 | 86.69 |
| 75% | 89.93 | 89.30 | 93.30 | 90.66 |
| 100% | **94.25** | **94.68** | **94.12** | **94.08** |

Table 4: Impact of Training Data Scale (%).

| Method | ACC | Precision | Recall | F1 |
|---|---|---|---|---|
| *JPEG Compression (Quality Factors = 25)* | | | | |
| R-MFDN | 75.68 | 83.85 | 62.80 | 71.82 |
| RAIRNet | 79.26 | 79.61 | 77.92 | 78.76 |
| GPT4o | 86.28 | 81.29 | 93.75 | 87.08 |
| Ours | **93.83** | **92.80** | **94.86** | **93.82** |
| *Gaussian Blur (kernel_size=(21, 21), sigma=8.0)* | | | | |
| RAIRNet | 59.87 | 59.33 | 59.27 | 59.30 |
| R-MFDN | 62.06 | 62.22 | 58.77 | 60.45 |
| GPT4o | 83.74 | 81.16 | 87.30 | 84.12 |
| Ours | **86.18** | **83.49** | **89.72** | **86.49** |

Table 5: Performance under perturbations (%)

## Performance under Perturbations

To evaluate the robustness of our method, we apply perturbation methods including JPEG compression and Gaussian Blur. The results in Table 5 demonstrate that our method exhibits robustness against these perturbations. Notably, even when Gaussian blur is applied to reduce low-level image information, our method maintains competitive performance, which can be attributed to the utilization of high-level semantic information in our approach.

# Conclusion

We present a personalized vision-language model for robust face forgery detection that integrates low-level artifact analysis with high-level semantic reasoning. Our approach uses specialized identifier tokens to encode identity-specific information from reference images, enabling effective detection without extensive fine-tuning. A lightweight detection adapter extracts low-level evidence while preserving the VLM's semantic capabilities. Training a new identity requires only modest computational overhead, as low as 100 images and 0.5 GPU hours for 10 tokens, making our approach highly scalable in real-world scenarios. Experiments on our IDImage dataset show our method outperforms traditional forgery detectors and VLMs with minimal additional parameters and reduced training data requirements.

## Acknowledgments

This work was supported by NSFC project (No. 62232006).

## References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security, WIFS 2018, Hong Kong, China, December 11-13, 2018*, 1–7. IEEE.

Alaluf, Y.; Richardson, E.; Tulyakov, S.; Aberman, K.; and Cohen-Or, D. 2024. MyVLM: Personalizing VLMs for User-Specific Queries. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XIII*, volume 15071 of *Lecture Notes in Computer Science*, 73–91. Springer.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *CoRR*, abs/2502.13923.

Cao, J.; Ma, C.; Yao, T.; Chen, S.; Ding, S.; and Yang, X. 2022. End-to-End Reconstruction-Classification Learning for Face Forgery Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 4103–4112. IEEE.

Chen, J.; Yao, J.; and Niu, L. 2024. A Single Simple Patch is All You Need for AI-generated Image Detection. *CoRR*, abs/2402.01123.

Chen, R.; Chen, X.; Ni, B.; and Ge, Y. 2020. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In *MM '20: The 28th ACM International Conference on Multimedia*.

Chollet, F. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 1800–1807. IEEE Computer Society.

Cozzolino, D.; Rössler, A.; Thies, J.; Nießner, M.; and Verdoliva, L. 2021. ID-Reveal: Identity-aware DeepFake Video Detection. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 15088–15097. IEEE.

Dong, X.; Bao, J.; Chen, D.; Zhang, T.; Zhang, W.; Yu, N.; Chen, D.; Wen, F.; and Guo, B. 2022. Protecting Celebrities from DeepFake with Identity Consistency Transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 9458–9468. IEEE.

Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Chen, D.; Wen, F.; and Guo, B. 2020. Identity-Driven DeepFake Detection. *CoRR*, abs/2012.03930.

Fang, M.; Yu, L.; Xie, H.; Wu, J.; Wang, Z.; Li, J.; and Zhang, Y. 2023. RAIRNet: Region-Aware Identity Rectification for Face Forgery Detection. In El-Saddik, A.; Mei, T.; Cucchiara, R.; Bertini, M.; Vallejo, D. P. T.; Atrey, P. K.; and Hossain, M. S., eds., *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, 1455–1464. ACM.

Foteinopoulou, N. M.; Ghorbel, E.; and Aouada, D. 2024. A Hitchhiker's Guide to Fine-Grained Face Forgery Detection Using Common Sense Reasoning. *Advances in Neural Information Processing Systems*, 37: 2943–2976.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2023. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Guo, Z.; Wu, Y.; Chen, Z.; Chen, L.; Zhang, P.; and He, Q. 2024. PuLID: Pure and Lightning ID Customization via Contrastive Alignment. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Huang, B.; Wang, Z.; Yang, J.; Ai, J.; Zou, Q.; Wang, Q.; and Ye, D. 2023. Implicit Identity Driven Deepfake Face Swapping Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 4490–4499. IEEE.

Huang, Z.; Xia, B.; Lin, Z.; Mou, Z.; Yang, W.; and Jia, J. 2024. Ffaa: Multimodal large language model based explainable open-world face forgery analysis assistant. *arXiv preprint arXiv:2408.10072*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; Madry, A.; Baker-Whitcomb, A.; Beutel, A.; Borzunov, A.; Carney, A.; Chow, A.; Kirillov, A.; Nichol, A.; Paino, A.; Renzin, A.; Passos, A. T.; Kirillov, A.; Christakis, A.; Conneau, A.; Kamali, A.; Jabri, A.; Moyer, A.; Tam, A.; Crookes, A.; Tootoonchian, A.; Kumar, A.; Vallone, A.; Karpathy, A.; Braunstein, A.; Cann, A.; Codispoti, A.; Galu, A.; Kondrich, A.; Tulloch, A.; Mishchenko, A.; Baek, A.; Jiang, A.; Pelisse, A.; Woodford, A.; Gosalia, A.; Dhar, A.; Pantuliano, A.; Nayak, A.; Oliver, A.; Zoph, B.; Ghorbani, B.; Leimberger, B.; Rossen, B.; Sokolowsky, B.; Wang, B.; Zweig, B.; Hoover, B.; Samic, B.; McGrew, B.; Spero, B.; Giertler, B.; Cheng, B.; Lightcap, B.; Walkin, B.; Quinn, B.; Guarraci, B.; Hsu, B.; Kellogg, B.; Eastman, B.; Lugaresi, C.; Wainwright, C. L.; Bassin, C.; Hudson, C.; Chu, C.; Nelson, C.; Li, C.; Shern, C. J.; Conger, C.; Barette, C.; Voss, C.; Ding, C.; Lu, C.; Zhang, C.; Beaumont, C.; Hallacy, C.; Koch, C.; Gibson, C.; Kim, C.; Choi, C.; McLeavey, C.; Hesse, C.; Fischer, C.; Winter, C.; Czarnecki, C.; Jarvis, C.; Wei, C.; Koumouzelis, C.; and Sherburn, D. 2024. GPT-4o System Card. *CoRR*, abs/2410.21276.

Jia, S.; Lyu, R.; Zhao, K.; Chen, Y.; Yan, Z.; Ju, Y.; Hu, C.; Li, X.; Wu, B.; and Lyu, S. 2024. Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, 4324–4333. IEEE.

Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.; and Shan, Y. 2024. PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 8640–8650. IEEE.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Nguyen, T.; Liu, H.; Li, Y.; Cai, M.; Ojha, U.; and Lee, Y. J. 2024. Yo'LLaVA: Your Personalized Language and Vision Assistant. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 22500–22510. IEEE.

s0md3v. 2023. Roop. https://github.com/s0md3v/roop.

Tan, C.; Liu, H.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024. Rethinking the Up-Sampling Operations in CNN-Based Generative Network for Generalizable Deepfake Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 28130–28139. IEEE.

Xu, J.; Chen, J.; Song, X.; Han, F.; Shan, H.; and Jiang, Y. 2024. Identity-Driven Multimedia Forgery Detection via Reference Assistance. In Cai, J.; Kankanhalli, M. S.; Prabhakaran, B.; Boll, S.; Subramanian, R.; Zheng, L.; Singh, V. K.; César, P.; Xie, L.; and Xu, D., eds., *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, 3887–3896. ACM.

Xu, Z.; Zhang, X.; Li, R.; Tang, Z.; Huang, Q.; and Zhang, J. 2025. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *ICLR*.

Yan, Z.; Zhang, Y.; Fan, Y.; and Wu, B. 2023. UCF: Uncovering Common Features for Generalizable Deepfake Detection. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 22355–22366. IEEE.

Zhang, Y.; Colman, B.; Guo, X.; Shahriyari, A.; and Bharaj, G. 2024. Common Sense Reasoning for Deepfake Detection. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXVIII*, volume 15146 of *Lecture Notes in Computer Science*, 399–415. Springer.

Zhou, Z.; Li, J.; Li, H.; Chen, N.; and Tang, X. 2024. StoryMaker: Towards Holistic Consistent Characters in Text-to-image Generation. *CoRR*, abs/2409.12576.