MEGRA: MIXTURE-OF-EXPERTS AND GRAPH RETRIEVAL-AUGMENTED GENERATION FOR EXPLAINABLE FACE FORGERY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conventional deepfake detection methods typically output only authenticity labels, forgery scores, or visual heatmaps, making it difficult to provide natural-language explanations that are readily understandable to humans. Recent methods based on multimodal large language models (MLLMs) can generate natural-language explanations, offering a new direction for explainable deepfake detection. However, they still face two challenges. First, their modeling of fine-grained forgery cues remains insufficient, resulting in poor generalization to forgery mechanisms unseen during training. Second, MLLM-generated explanations are prone to hallucination. To mitigate these issues, we propose MEGRA (Mixture-of-Experts and Graph Retrieval-Augmented Generation), a face forgery detection method that combines mechanism-aware Mixture-of-Experts (MoE) detection with Graph Retrieval-Augmented Generation (GraphRAG) explanation. MEGRA comprises two modules: a forgery detection module and an explanation module. In the forgery detection module, we design a mechanism-oriented Forgery Detection MoE (FD-MoE) that adaptively combines experts with complementary forensic capabilities to better discriminate among different forgery mechanisms and adapt to distribution shifts. In the explanation module, we construct face-Forensic Knowledge Graphs (F-FKGs) and introduce GraphRAG to retrieve relevant forensic knowledge conditioned on the detection result and regional forensic evidence, guiding the MLLM to generate evidence-grounded natural-language explanations. Experiments on HydraFake show that MEGRA achieves an average accuracy of 91.09%, reaching state-of-the-art (SOTA) performance. In addition, MEGRA's explanations receive the highest scores among reported results under all three judge models: GPT-5.5, Kimi-K3, and Qwen3.7-Plus.

1 INTRODUCTION

Rapid advances in generative artificial intelligence (AI) continue to improve face synthesis and editing, enabling facial reenactment and related media-production applications to approach photorealistic quality (Thies et al., 2016) while substantially increasing the realism of deepfake content (Rössler et al., 2019; Tan et al., 2026). As these technologies become widely accessible, the risks of impersonation, fraud, and disinformation increase, making reliable face forgery detection an important research problem. Meanwhile, deepfake generation has expanded beyond conventional face swapping and reenactment to diverse mechanisms such as diffusion-based synthesis, personalization, and relighting. This diversity places greater demands on a detector's ability to generalize to generation mechanisms and data domains unseen during training (Tan et al., 2026). As deepfake detection is increasingly applied to identity verification and content-authenticity assessment, the need for trustworthy and interpretable detection results is also growing.

In response, deepfake detection has increasingly shifted from authenticity classification alone toward jointly addressing detection and explanation (Guo et al., 2025). Recent MLLM-based methods exploit visual understanding and language generation to produce natural-language explanations in addition to authenticity judgments (Huang et al., 2024; Tan et al., 2026). However, existing methods still face two important challenges: detection generalization and explanation reliability. First, their limited modeling of fine-grained forgery cues restricts generalization to unseen forgery mech-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

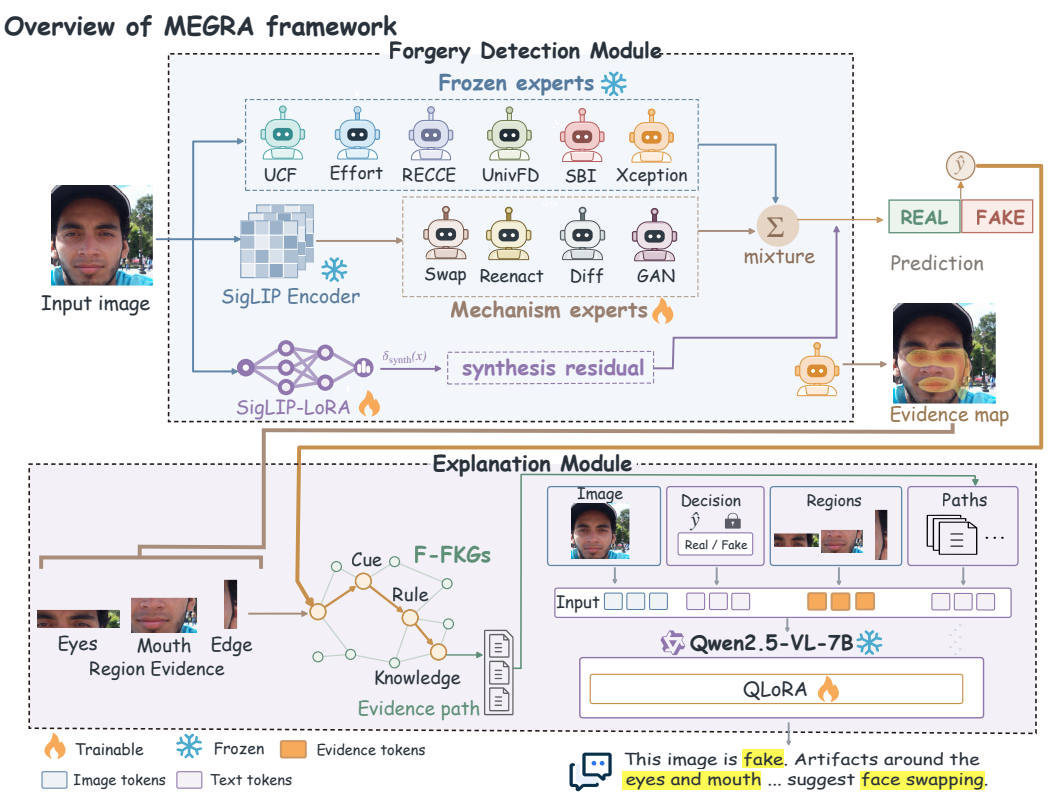


Figure 1: Visual content is for illustration only and is not drawn from evaluation samples.

anisms and data domains (Tan et al., 2026). Second, open-ended MLLM generation can hallucinate, causing explanations to conflict with the actual forensic evidence (Li et al., 2023). Improving cross-mechanism and cross-domain detection while grounding generated explanations in reliable forensic evidence therefore remains a central problem in explainable deepfake detection.

To this end, we propose MEGRA, a face forgery detection method combining mechanism-aware detection and knowledge-enhanced explanation. MEGRA jointly designs forgery detection and explanation generation: the former improves generalization across diverse forgery mechanisms and distribution shifts, and the latter introduces forensic evidence and structured knowledge to support MLLM explanation, thereby addressing both detection performance and explanation quality.

The main contributions of this paper are:

1. We propose MEGRA, an explainable deepfake face detection method combining mechanism-aware MoE and GraphRAG.
2. In MEGRA's forgery detection module, we design FD-MoE to integrate heterogeneous experts with complementary forensic capabilities according to latent forgery mechanisms, improving generalization to unseen forgery mechanisms and data domains.
3. In MEGRA's explanation module, we construct F-FKGs and introduce GraphRAG into explainable deepfake face detection, enabling the MLLM to generate explanations based on the detection result, regional forensic evidence, and structured forensic knowledge.
4. Experiments show that MEGRA achieves 91.09% average accuracy on HydraFake (Tan et al., 2026), reaching SOTA performance; its explanations score highest among reported results under three judges: GPT-5.5, Kimi-K3, and Qwen3.7-Plus.

2 Related Work

So far, deepfake face detection has attracted considerable attention. Existing research mainly addresses forgery cue modeling and explainable detection and can be divided into conventional visual

detection methods and explainable methods based on large language models (LLMs). This section reviews these two types of methods and further discusses MoE and retrieval-augmented generation (RAG) techniques relevant to MEGRA.

2.1 DEEPFAKE DETECTION

Conventional deepfake detectors use CNNs or Transformers to model forensic cues such as visual artifacts, blending boundaries, frequency patterns, and identity consistency (Li et al., 2020; Qian et al., 2020; Dong et al., 2022; Nadimpalli & Rattani, 2023). Subsequent work improves generalization to unseen forgery methods and data domains through feature disentanglement and pretrained visual representations (Yan et al., 2023a; Ojha et al., 2023; Yan et al., 2025). These methods primarily target authenticity classification and cannot readily provide natural-language explanations. Recent MLLM-based deepfake detectors have begun to address authenticity judgment and natural-language explanation jointly. M2F2-Det combines visual detection with an LLM (Guo et al., 2025); FFAA formulates open-world face forgery analysis as visual question answering (Huang et al., 2024); Yu et al. use detection and localization outputs to guide language generation (Yu et al., 2025); and $\mathcal{X}^2$-DFD supplements weak MLLM forensic features with specialized detectors (Chen et al., 2025). VERITAS further improves adaptation to unseen forgeries and domains through pattern-aware reasoning (Tan et al., 2026). In broader image forensics, FakeShield, SIDA, and FakeVLM also explore detection and explanation (Xu et al., 2025; Huang et al., 2025; Wen et al., 2025). Nevertheless, existing methods remain limited in both detection and explanation: insufficient modeling of fine-grained forgery cues restricts generalization to unseen forgery mechanisms and data domains (Tan et al., 2026), while open-ended MLLM generation can hallucinate, producing explanations inconsistent with the actual forensic evidence (Li et al., 2023).

2.2 MOE AND RAG

MoE and knowledge-enhanced generation techniques have been introduced into visual understanding and multimodal reasoning to improve complex-pattern modeling and the reliability of generated results. This subsection reviews MoE and RAG techniques relevant to MEGRA.

MoE. The original MoE formulation assigns weights to multiple specialized experts according to the input through a gating mechanism and combines their predictions; it was later extended to hierarchical expert structures (Jacobs et al., 1991; Jordan & Jacobs, 1994). With the development of LLM, sparse MoE has been widely used to scale model capacity by activating only a subset of the internal subnetworks (Shazeer et al., 2017; Fedus et al., 2022). Our work focuses on the functional specialization of forensic experts, assigning different forensic tasks to independent experts and combining them through a gating mechanism. This design is closer to the expert-collaboration paradigm of early MoE. In deepfake detection, MoE-FFD places gated low-rank adaptation (LoRA) (Hu et al., 2022) and Adapter experts in a frozen vision Transformer (ViT) to combine global representations and local forgery priors (Kong et al., 2026); MEGRA's FD-MoE uses an internal mechanism-aware mixture network to adaptively assign fusion weights between frozen forensic experts and mechanism experts, integrating complementary forensic capabilities and inductive biases.

RAG. RAG enhances language generation by retrieving relevant evidence from an external corpus (Lewis et al., 2020). In explainable deepfake detection, RAIDX incorporates retrieved information from related images into reasoning (Li et al., 2025), while VRAG-DFD retrieves forensic knowledge to enhance MLLM reasoning (Han et al., 2026). However, similarity-based retrieval typically treats evidence as relatively independent text fragments and cannot explicitly model the relations among forgery mechanisms, local regions, and forensic cues. GraphRAG instead organizes candidate evidence through entity relations (Edge et al., 2024) and retrieves explicitly connected knowledge paths along the graph (Sun et al., 2024), providing more structured and traceable evidence for generation. Nguyen & Stamm (2026) further introduce Forensic Knowledge Graphs (FKGs) for general image authentication, organizing dynamically identified forensic regions, traces, and their dependencies within a single image into an image-specific forensic graph. In contrast, we focus on deepfake face detection and build shared F-FKGs. Given the detection result and aligned regional evidence, GraphRAG retrieves relevant forensic paths to support MLLM explanations.

3 METHOD

3.1 PROBLEM SETUP

Given a face image x with label $y \in \{0, 1\}$ (0: real; 1: fake), the detection module outputs a score $S(x)$ and decision $\hat{y} = \mathbf{1}[S(x) \geq \tau]$, where τ is fitted only on the validation set. For a set of N samples $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, the detection objective is

$$\max_{\theta_{\text{det}}} \text{Acc}(\mathcal{D}; \theta_{\text{det}}) = \max_{\theta_{\text{det}}} \frac{1}{N} \sum\nolimits_{i=1}^{N} \mathbf{1}[\hat{y}_i(\theta_{\text{det}}) = y_i]. \tag{1}$$

This objective also evaluates the model's generalization to unseen forgery mechanisms and different data domains. The explanation task generates a natural-language explanation Y_x by optimizing

$$\max_{\theta_{\text{exp}}} \text{Score}(Y_x; x, \hat{y}), \quad \text{s.t. Decision}(Y_x) = \hat{y}. \tag{2}$$

Here, $\text{Score}(Y_x; x, \hat{y})$ measures whether the explanation is consistent with the detection result $\hat{y}$ and faithfully describes the forgery cues in x and their supporting evidence.

3.2 MEGRA

MEGRA comprises a forgery detection module and an explanation module (Figure 1). The forgery detection module outputs an authenticity score and decision and extracts regional forensic evidence for explanation. Conditioned on the fixed detection decision and regional evidence, the explanation module introduces structured forensic knowledge through GraphRAG and generates a natural-language explanation. The designs of the two modules are described below.

3.2.1 FORGERY DETECTION MODULE

The forgery detection module comprises FD-MoE, a synthesis residual, and a regional evidence side branch. FD-MoE is the core detector and includes frozen forensic experts, mechanism experts, expert calibration, a detector-side SigLIP-LoRA channel, and a mechanism-aware mixture network. The two types of experts provide generic forensic priors and mechanism-specific evidence, respectively. The mechanism-aware mixture network estimates expert combination weights according to the current sample's latent forgery mechanism, yielding the core detection score $S_{\text{FD-MoE}}(x)$. The synthesis residual $\Delta(x)$ adds entire-face-synthesis evidence to obtain the final score $S(x) = S_{\text{FD-MoE}}(x) + \Delta(x)$.

Frozen forensic experts. Let $\mathcal{E}_{\text{f}} = \{E_e\}_{e=1}^{K}$ denote the set of frozen forensic experts, where K is the number of experts. For an input image x, each expert E_e outputs a forensic logit $z_e(x) = E_e(x)$. The experts address complementary forensic cues and remain frozen during training of the detection module, jointly providing generic forensic priors.[1] Because the polarity and scale of expert scores differ, we calibrate the polarity, temperature, and bias of logit $z_e(x)$ for each mechanism class c:

$$s_{e,c}(x) = Q_e\left(\sigma\left(\frac{a_{e,c} z_e(x) - b_{e,c}}{T_{e,c}}\right)\right), \tag{3}$$

where $a_{e,c} \in \{\pm 1\}$, $T_{e,c} > 0$, and $b_{e,c}$ are fitted only on the validation set, while the monotonic quantile map Q_e is shared across classes. Details are provided in Appendices C and B.1.

Mechanism experts. Frozen experts provide generic priors but cannot fully cover mechanism-specific traces in face swapping, reenactment, and entire-face synthesis. We therefore build four MLP experts on the frozen SigLIP embedding $\phi(x)$ (Zhai et al., 2023) to supplement evidence specific to $c \in \mathcal{C} = \{\text{Swap}, \text{Reenact}, \text{Diff}, \text{GAN}\}$:

$$d_c(\phi(x)) = \sigma\big(W_{c,2}\, \text{ReLU}(W_{c,1}\phi(x) + b_{c,1}) + b_{c,2}\big). \tag{4}$$

Each binary expert is trained on fake images from its corresponding mechanism and all real images. Swap/Reenact emphasize local manipulation, whereas Diff/GAN emphasize entire-face generation.

[1] In our experiments, six of eight candidate frozen experts form the final expert pool; the six-expert configuration achieves the best overall performance. Candidate expert selection and analyses of expert pools of different sizes are provided in Appendices C and J and Table 18.

Mechanism-aware mixture network. As an internal fusion component of FD-MoE, the mechanism-aware mixture network is designed to avoid fixed fusion that favors mechanisms dominant in training. All expert channels run in parallel, and the mixture network adaptively adjusts their contributions according to the current image's latent mechanism. Let $\mathcal{H}$ denote the frozen experts, mechanism experts, and detector-side synthesis LoRA channel. Class-conditional nonnegative weights $w_{c,h}$ define a mechanism margin

$$m_c(x) = \sum\nolimits_{h\in\mathcal{H}} w_{c,h}\, s_{h,c}(x) - \tau_c - \delta_c, \tag{5}$$

where τ_c and δ_c are the class-specific threshold and offset. The weights w_c are fitted to maximize balanced accuracy (BA) on leave-one-forgery-method-out validation sets. To estimate sample-level mixture coefficients, we group Swap/Reenact into the local-manipulation family and Diff/GAN into the entire-face-synthesis family. A factorized head then estimates from visual features

$$p(c \mid x) = p(f(c) \mid x)\, p(c \mid f(c), x). \tag{6}$$

Combining this distribution with the expert-response distribution $\pi(c \mid x)$ and the mechanism margins yields adaptive coefficients $\tilde{p} = \mathrm{Norm}(\mathrm{Refine}(p, \pi, m))$, with refinement and training objectives detailed in Appendix B.1. The FD-MoE detection score is $S_{\text{FD-MoE}}(x) = \sum_c \tilde{p}_c(x)\, m_c(x)$. Thus, the contribution of the same forensic channel can vary with the sample's latent mechanism, avoiding detection rules fixed to a particular seen forgery type.

Synthesis residual. Although FD-MoE adaptively integrates evidence, it may still underuse evidence for some entire-face synthetic images and assign them excessively high real confidence, i.e., judge fake samples as "too real." Let $c = \arg\max_{c'} \tilde{p}_{c'}(x)$ and $F_c = \sum_{h\in\mathcal{H}} w_{c,h} s_{h,c}(x)$. Two class-calibrated SigLIP-LoRA channels output a synthesis score $s_{\text{synth},c}$ and an identity-preserving score $s_{\text{ID},c}$. For entire-face-synthesis samples with an excessively low forgery score, a one-sided residual increases the score when additional evidence exceeds the current fused evidence; for local-manipulation samples, a separate branch activates when diffusion-class synthesis evidence exceeds the fused evidence:

$$\begin{aligned}\Delta(x) = \mathbf{1}_{\text{synth}}(\alpha_{\text{ID}}[s_{\text{ID},c} - F_c]_+ + \alpha_{\text{synth}}[s_{\text{synth},c} - F_c]_+) \\ + \mathbf{1}_{\text{photo}}\alpha_{\text{synth}}[s_{\text{synth,Diff}} - F_c]_+.\end{aligned} \tag{7}$$

Here, $[u]_+ = \max(0, u)$, and $\mathbf{1}_{\text{synth}}$ and $\mathbf{1}_{\text{photo}}$ indicate compensation conditions for entire-face synthesis and local manipulation, respectively. Detailed information is provided in Appendix B.

Regional evidence side branch. During the same forward pass, a frozen Xception evidence side branch extracts evidence R_x from eight semantically aligned facial regions. It is used only for explanation and does not contribute to the computation of $S(x)$.

3.2.2 Explanation Module

The explanation module does not change the detection decision; it explains the basis for the prediction using the fixed decision, regional evidence from the current image, and structured forensic knowledge. The module first constructs F-FKGs shared across samples, then retrieves sample-relevant forensic paths through GraphRAG and uses them for supervised fine-tuning (SFT) and explanation generation. This organizes the fixed detection decision and regional evidence into a structured basis for explanation, improving decision consistency, sample relevance, and traceability while reducing the risk of inconsistency between open-ended generation and actual forensic evidence. Specifically, the explanation module queries F-FKGs $G = (V, A)$ with $(\hat{y}, R_x)$ for sample-relevant typed evidence paths $\mathcal{P}_x$ and generates Y_x conditioned on $(x, \hat{y}, R_x, \mathcal{P}_x)$. The authenticity field is locked to the upstream decision by Eq. (2), ensuring detection–explanation consistency.

F-FKGs. Existing forensic knowledge is commonly stored as isolated text fragments, making it difficult to explicitly model dependencies among facial regions, anomalous cues, decision rules, and forgery types, or to trace the knowledge supporting an explanation. To address this limitation, we construct F-FKGs shared across samples and organize distributed forensic knowledge into typed region–cue–rule–type–text structures. Region nodes provide local anchors, cue nodes represent candidate anomalies, and rule nodes combine regions with cues and connect them to compatible forgery types and supporting text. Beyond forgery cues, the graph includes authenticity patterns such as global consistency, texture–geometry coherence, and illumination–context coherence. It also

contains caution rules for ambiguous cases involving compression, blur, or low resolution to limit overstatement. These nodes are connected by typed relations such as localized-at, associated-with, supports, described-by, and requires-caution. The detailed composition is given in Appendix D. Generator names, data splits, sample identifiers, and sample-level ground truth are excluded during graph construction to prevent dataset metadata from being used as forensic evidence.

GraphRAG. Based on F-FKGs, GraphRAG converts detection and regional information for the current sample into evidence paths relevant to explanation generation. GraphRAG retrieves sample-specific evidence from the graph conditioned on the detection decision $\hat{y}$, high-response regions R_x, the coarse mechanism prior $\arg\max_c \tilde{p}_c$, and uncertainty. Let r, u, t, and d denote region, cue, type, and text, respectively. For a fake decision, each candidate evidence path is a typed star subgraph

$$P = (V_P, E_P), \qquad V_P = \{r, u, \eta, t, d\}, \quad E_P = \{(r, \eta), (u, \eta), (\eta, t), (\eta, d)\}, \tag{8}$$

where the rule η connects r, u, t, and d. For a real decision, the system instead retrieves region–authenticity-pattern–text paths; for highly uncertain samples, the results may also include a caution strategy. Candidate paths must overlap the current regions and candidate cues and are ranked by

$$\rho(P) = \lambda^\top q(P; R_x, C_x, \hat{y}, U_x), \tag{9}$$

where q jointly measures decision compatibility, region/cue overlap, evidence completeness, uncertainty, and evidence strength. We retain at most five paths of up to four hops each; graph and retrieval settings are detailed in Appendix D.

SFT. The retrieved evidence paths serve as external context during SFT. A QLoRA-adapted Qwen2.5-VL-7B (Bai et al., 2025; Dettmers et al., 2023) takes the image, fixed decision, regional evidence, and path text as inputs, with residual evidence injected through projected global and regional descriptors. Training uses masked token-weighted SFT so that the MLLM learns to organize forensic explanations using the retrieved knowledge. For a minibatch $\mathcal{B}_s$, let $\mathcal{I}_i = (x_i, \hat{y}_i, R_{x_i}, \mathcal{P}_{x_i})$ denote the conditioning input and $u_{i,t}$ the t-th token of the supervised explanation. The actual optimization objective is

$$\mathcal{L}_{\text{exp}}(\mathcal{B}_s) = -\frac{\sum_{i \in \mathcal{B}_s} \sum_t m_{i,t} w_{i,t} \log \pi_\theta(u_{i,t} \mid \mathcal{I}_i, u_{i,<t})}{\sum_{i \in \mathcal{B}_s} \sum_t m_{i,t} w_{i,t}}. \tag{10}$$

Here, $m_{i,t}$ masks padding and prompt tokens, while $w_{i,t}$ downweights JSON-structure tokens and keeps explanation-content tokens at weight 1. Teacher targets include both positive forensic examples and negative phrasing constraints, a design inspired by contrastive in-context learning (Gao & Das, 2024). Weighting, training, and residual-injection details are provided in Appendix E. At inference, the system retrieves evidence paths for the current sample as input to the fine-tuned MLLM. For fake decisions, the model generates explanations such as face swapping, reenactment, or entire-face synthesis. We check for decision reversals, unsupported region/cue references, and prohibited metadata, and fix the final decision field to $\hat{y}$ according to Eq. (2).

4 Experiments

4.1 Experimental Setup

We use HydraFake's official train/validation/test split; its test protocols comprise in-domain (ID; approximately 14k images), cross-model (CM; 11k), cross-forgery (CF; 12k), and cross-domain (CD; 15k) subsets (Tan et al., 2026) (Appendix A). CM and CF respectively evaluate adaptation to unseen generation models and unseen manipulation methods. CD further includes forgery methods and forgery types unseen during training, together with unseen real-image domains, making it our key setting for measuring open-world generalization. Following prior work (Tan et al., 2026), we use accuracy as the primary detection metric and additionally report area under the curve (AUC) and other metrics in Appendix I. To reduce variation caused by different reimplementation environments, all non-MEGRA results in Table 1 are taken directly from VERITAS, while MEGRA is evaluated on the corresponding test subsets. The FaceVid2Vid subset has different class composition for MEGRA. The explanation task is evaluated using Score Evaluation and pairwise ELO rating. To ensure consistent and fair evaluation, we use the public evaluation prompts and scoring

Table 1: Accuracy (%) under HydraFake's stratified ID/CM/CF/CD test protocol. Avg. is the unweighted mean over ID and the 18 OOD subsets; for rows with †ID shown as "–", Avg. is computed over the 18 OOD subsets only. Published detectors are annotated with venue and year; cd denotes the cold-start variant of VERITAS. **Bold**/underline: best/second-best result in each column among methods reporting Avg. in the same form, including ties.

		Cross-Model						Cross-Forgery						Cross-Domain						
Method	ID	*ADF*	*FLUX*	*StAI*	*MAGI*	*HART*	*Inf.*	*StGv2*	*ICL*	*CodeF*	*InfYou*	*PuLID*	*FcAda*	*DFL*	*IY-CD*	*Dream*	*Hailuo*	*G4o*	*FFIW*	Avg.
Compact vision detectors																				
F3Net (ECCV'20)	85.3	86.7	87.8	78.6	85.0	86.0	82.9	41.3	48.9	71.9	84.9	85.5	72.6	57.7	78.5	55.6	68.6	66.2	66.4	73.2
UnivFD (CVPR'23)	82.7	90.7	93.8	82.5	73.0	94.4	90.7	61.8	81.9	75.4	73.7	68.1	81.3	67.4	67.3	80.5	75.2	73.3	67.5	78.0
IID (CVPR'23)	83.4	83.3	82.8	80.0	80.2	81.1	82.2	41.4	53.3	79.7	81.8	81.8	73.7	65.2	69.9	63.8	63.3	63.8	64.2	72.4
FreqNet (AAAI'24)	66.8	60.3	76.7	59.0	69.2	77.1	75.1	33.1	73.1	70.3	72.8	77.4	67.7	50.6	67.0	62.1	59.3	58.3	51.2	64.6
ProDet (NeurIPS'24)	90.5	92.6	94.2	88.2	91.9	93.8	93.1	56.3	58.6	80.8	88.1	91.0	83.3	58.1	82.9	71.3	75.6	66.3	74.1	80.6
NPR (CVPR'24)	75.6	68.8	91.2	59.5	82.6	91.3	84.0	47.7	67.8	60.6	79.8	89.0	67.7	52.6	73.0	76.6	62.3	50.2	46.0	69.8
AIDE (ICLR'25)	80.4	68.8	86.3	64.0	88.9	95.4	76.0	56.7	79.2	86.1	74.2	62.4	75.7	59.7	67.9	49.7	58.0	51.9	59.2	70.6
Co-SPY (CVPR'25)	86.3	93.5	95.5	85.3	93.3	96.6	95.3	77.0	92.5	88.6	90.6	79.1	87.3	67.6	80.0	82.5	74.0	79.5	64.3	84.7
D^3 (CVPR'25)	87.3	93.6	95.6	91.3	90.7	95.8	95.5	62.4	71.6	82.9	80.0	82.4	73.7	69.7	74.6	78.1	70.9	80.8	64.3	81.1
Effort (ICML'25)	94.7	82.8	96.5	78.0	90.5	97.8	98.3	64.7	**94.8**	89.7	89.5	92.9	88.0	64.8	82.2	61.5	66.4	53.8	74.0	82.2
Generic MLLMs																				
Qwen2.5-VL-7B	51.2	50.0	50.0	49.7	50.0	52.0	52.9	50.5	56.7	50.7	53.6	54.5	51.6	50.7	53.6	80.2	67.5	52.5	50.5	54.1
InternVL3-8B	54.0	54.0	49.8	49.0	56.6	55.8	57.2	62.9	54.2	62.9	63.6	54.8	67.7	54.4	67.1	77.1	66.5	47.4	51.8	58.3
MiMo-VL-7B	63.8	74.5	77.1	82.5	60.3	82.4	81.4	48.7	82.6	76.4	79.7	78.4	82.8	57.7	75.6	79.4	70.7	67.7	54.9	72.5
GLM-4.1V-T	56.4	55.2	52.3	50.5	51.6	68.4	60.7	54.3	68.4	63.3	65.7	55.1	81.0	58.7	72.7	83.7	69.2	52.0	53.9	61.7
GPT-4o	53.5	57.7	52.0	51.4	59.9	81.2	54.8	66.4	58.9	52.5	64.4	60.9	55.5	49.4	62.0	90.7	73.7	58.0	52.8	60.8
Gemini-2.5-Pro	72.2	64.9	92.4	82.8	62.5	93.4	93.2	73.7	83.3	87.4	85.5	84.7	85.6	67.2	75.6	87.5	82.4	70.9	53.0	78.9
MLLM-based forgery detectors																				
M2F2-Det (CVPR'25)	–†	56.0	57.7	59.8	61.8	61.3	55.4	78.9	65.5	80.0	57.4	57.5	76.3	73.0	56.3	67.2	50.6	53.0	70.6	63.2 †
FakeShield (ICLR'25)	–†	64.3	64.0	61.5	63.1	61.8	63.3	64.0	57.3	60.9	58.1	63.6	63.7	50.2	83.8	53.8	51.3	53.9	55.6	60.8 †
SIDA-7B (CVPR'25)	–†	97.3	97.7	79.5	59.3	98.5	95.0	59.8	60.6	62.3	89.7	94.4	63.3	50.4	81.9	80.0	78.0	68.9	57.3	76.3 †
SIDA-13B (CVPR'25)	–†	80.7	78.5	54.8	52.5	91.3	82.4	63.7	61.2	68.2	56.7	67.1	84.3	60.8	58.2	88.3	74.0	74.1	59.9	69.8 †
FFAA (arXiv'24)	–†	55.1	50.9	72.9	63.5	60.8	57.6	82.7	70.9	71.8	58.4	62.4	86.0	67.7	58.4	55.3	59.2	49.6	68.3	64.0 †
FakeVLM (NeurIPS'25)	–†	78.2	78.5	77.0	74.5	76.5	76.8	70.8	76.2	76.2	76.9	76.5	77.7	75.7	83.6	81.5	80.8	78.7	74.5	77.3 †
VERITAS-MINI (ICLR'26)	–†	95.5	99.1	97.3	72.8	97.0	96.1	82.5	76.3	90.0	83.7	82.9	79.3	72.5	78.7	92.0	93.0	85.5	70.6	85.8 †
VERITAS (cd) (ICLR'26)	96.8	79.5	99.6	96.0	**99.9**	99.7	**99.9**	84.0	65.3	94.8	86.2	93.4	86.7	55.9	73.5	93.7	89.3	88.1	76.4	87.3
VERITAS (ICLR'26)	**97.3**	94.8	**99.8**	**97.0**	**99.9**	**99.9**	**99.9**	90.3	75.7	**97.0**	**91.8**	**95.1**	91.7	58.6	**84.1**	92.3	90.2	89.2	**78.5**	90.7
MEGRA	95.5	**97.5**	97.3	91.5	98.1	97.7	98.0	**91.4**	84.4	92.3	89.4	89.9	**93.2**	**80.1**	80.6	**96.1**	**91.2**	**93.2**	73.3	**91.09**

ADF=Adobe Firefly; StAI=StarryAI; MAGI=MAGI-1; Inf.=Infinity; StGv2=StarGANv2; ICL=IC-Light; CodeF=CodeFormer; InfYou=InfiniteYou; FcAda=FaceAdapter; DFL=DeepFaceLab; IY-CD=InfiniteYou-CD; Dream=Dreamina; Hailuo=Hailuo AI; G4o=GPT-4o.

criteria from VERITAS (Tan et al., 2026). Specifically, we randomly sample 1,000 images (300 real and 700 fake): GPT-5.5, Kimi-K3, and Qwen3.7-Plus assign Score Evaluation ratings from 1 to 5, while Kimi-K3 and Qwen3.7-Plus conduct pairwise comparisons. The complete judging procedure, prompts, and ELO conversion formulas are provided in Appendices K and L.

4.2 Forgery Detection Performance

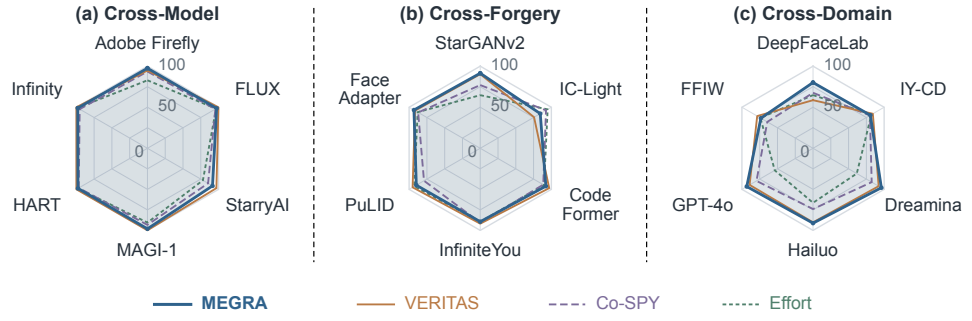


Figure 2: OOD accuracy on the CM, CF, and CD splits using a shared scale and Table 1.

As shown in Table 1, MEGRA achieves an average accuracy of 91.09% (95% confidence interval: $[90.77, 91.42]$), exceeding VERITAS at 90.7%. This advantage mainly comes from the OOD subsets, where average accuracy increases from 90.32% to 90.84%. On the most challenging CD subsets, which more closely resemble real-world applications, average accuracy improves from 82.15% to 85.74%. Because CD jointly covers forgery methods, forgery types, and real-image domains unseen during training, this result indicates that MEGRA generalizes more effectively to distribution shifts in open-world settings. Specifically, MEGRA achieves substantial gains on DeepFaceLab (80.1% vs. 58.6%), IC-Light (84.4% vs. 75.7%), GPT-4o (93.2% vs. 89.2%), and Dreamina (96.1% vs. 92.3%), demonstrating adaptation across generation mechanisms and data domains. Figure 2 further shows that MEGRA has a more balanced accuracy profile across OOD subsets than VERITAS, Co-SPY, and Effort; its standard deviation over the 18 OOD subsets is 6.88 percentage points, compared with 10.29 for VERITAS. Results on ID (95.5% vs. 97.3%) and FFIW (73.3% vs. 78.5%) further indicate that the overall improvement is driven mainly by stronger performance on cross-domain subsets, with further gains on individual unseen generators and manipulation methods.

4.3 Explanation and Reasoning Evaluation

Table 2: Score Evaluation (1–5) ↑ and pairwise ELO rating ↑ on 1,000 randomly sampled images. Venue tags follow Table 1. Best results are in **bold**.

	Score Evaluation ↑			ELO Rating ↑	
Method	GPT-5.5	Kimi-K3	Qwen3.7-Plus	Kimi-K3	Qwen3.7-Plus
Qwen2.5-VL-7B (base)	–	1.63	1.97	764.8	–
M2F2-Det (CVPR'25)	–	1.55	1.77	644.1	895.5
SIDA-7B (CVPR'25)	–	1.62	1.59	402.8	762.6
FakeVLM (NeurIPS'25)	–	2.95	2.33	906.6	744.0
VERITAS (cd) (ICLR'26)	–	3.58	4.21	1271.4	1045.7
VERITAS (ICLR'26)	3.78	3.59	4.30	1287.0	1043.5
MEGRA	**4.09**	**4.18**	**4.75**	**1723.3**	**1438.5**

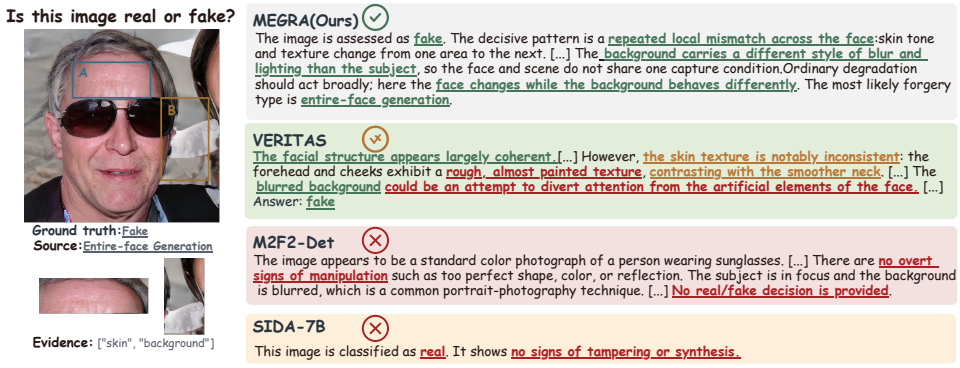


Figure 3: Qualitative comparison on one entire-face synthesis example from seeprettyface.

Table 2 shows that MEGRA achieves the highest Score Evaluation under GPT-5.5, Kimi-K3, and Qwen3.7-Plus, and the highest pairwise ELO under Kimi-K3 and Qwen3.7-Plus. Thus, the fixed detection decision, regional evidence, and GraphRAG-retrieved knowledge jointly improve the quality of the evidence used in explanations. Figure 3 shows one seeprettyface entire-face synthesis example. MEGRA uses the skin and background regions as evidence, relates local facial texture changes to differences in blur and lighting style between the subject and background, and attributes the image to entire-face synthesis after considering possible non-forgery factors such as compression, defocus, and denoising. VERITAS produces the correct fake label and identifies an anomaly in skin texture, but further interprets background blur as a generative strategy for diverting attention, without verifiable regional support. M2F2-Det only provides a general image description without an authenticity decision; SIDA-7B incorrectly classifies the example as real. Appendix N provides 20 further qualitative examples and two cases with correct fake decisions but explanation errors.

4.4 ABLATIONS

Table 3 presents detection variants based on the main components of the forgery detection module in Figure 1(a). Each variant retains the same frozen checkpoints, disables, zeros, or replaces only the target component, and refits the threshold using validation BA. Detailed definitions and frozen expert-pool baselines are provided in Appendices B.2 and J.

Table 3: Detection-module ablations (%). Each lesion refits the validation BA threshold. ID/CM/CF/CD and Avg. follow Table 1; **bold** denotes full MEGRA.

Setting	ID	CM	CF	CD	Avg.
MEGRA	**95.52**	**96.69**	**90.10**	**85.74**	**91.09**
w/o frozen experts	91.13	96.29	85.69	79.94	87.51
w/o mechanism experts	82.23	86.22	84.89	84.49	85.04
w/o class calib.	93.04	95.51	87.87	82.44	88.84
Max-margin	94.60	96.04	88.51	84.45	89.93
w/o residual	94.31	94.24	88.29	83.31	88.91

The detection ablations confirm the complementary roles of the components. Removing the mechanism experts causes the largest drop in average accuracy (6.05 percentage points), with the most pronounced degradation on ID and CM. Removing the frozen forensic experts reduces average accuracy by 3.58 points and primarily affects CF and CD. This suggests that mechanism experts emphasize representations specific to the defined mechanisms, whereas frozen experts provide generic forensic priors for distribution shifts. Without class-conditional calibration, average accuracy drops by 2.25 points, with the largest degradation on CD, indicating that aligning the polarity and scale of heterogeneous expert outputs benefits cross-domain mixture. Replacing the adaptive mixture with Max-margin reduces accuracy under all four protocols and lowers the average by 1.16 points; removing the synthesis residual causes a 2.18-point drop. These results support the contributions of sample-level expert mixture and entire-face-synthesis evidence compensation. Paired statistical significance tests are reported in Appendix H.

Table 4: Explanation-module ablations using one fixed SFT checkpoint. ELO uses a separate three-method pool from Table 2; **bold** denotes full MEGRA.

	Score Evaluation ↑			ELO Rating ↑	
Method	GPT-5.5	Kimi-K3	Qwen3.7-Plus	Kimi-K3	Qwen3.7-Plus
MEGRA	**4.09**	**4.18**	**4.75**	**1126.7**	**1156.8**
w/o F-FKGs	–	4.08	4.48	972.8	916.2
w/o F-FKGs and regions	–	4.12	4.57	900.4	927.0

In Table 4, the full model performs best under both Score Evaluation metrics and both ELO metrics. Removing F-FKGs lowers the Kimi-K3/Qwen3.7-Plus scores from 4.18/4.75 to 4.08/4.48. Removing both F-FKGs and regional evidence yields a slight recovery to 4.12/4.57, although ELO does not show a consistent rebound (900.4/927.0). This non-monotonic behavior may arise because regional responses primarily provide localization: without structured forensic knowledge, the model may over-rely on local responses; removing both inputs leads the model to rely on forensic patterns internalized during SFT, producing some recovery in Score Evaluation. The full model nevertheless remains best on every metric, indicating that regional evidence and GraphRAG are complementary through localization and knowledge constraints.

5 CONCLUSION

We propose MEGRA, a face forgery detection method combining mechanism-aware FD-MoE detection and GraphRAG explanation. It achieves SOTA performances in terms of generalization to unseen forgeries through complementary forensic experts and uses structured forensic knowledge to guide the MLLM in generating evidence-grounded natural-language explanations.

ETHICS STATEMENT

This work studies synthetic-face detection and explanation on HydraFake and its cited sources; the knowledge graph, prompts and explanations avoid identity targeting and test-set generator identifiers, and we will not release attack recipes.

REPRODUCIBILITY STATEMENT

Dataset and evaluation protocols are described in §4.1 and Appendix A; implementation details and training settings are in Appendices B–E; training and inference cost relative to VERITAS is in Appendix F; hyperparameter sensitivity and paired tests of detection-module lesions are in Appendices G–H; additional metrics and the frozen expert pool are in Appendices I–J; prompts and judge procedures are in Appendices K–L; and additional qualitative examples are in Appendix N. The anonymous supplementary archive accompanying Appendix M contains a README, environment specifications, configuration files, data-processing notes, split indices, category mappings and the average-accuracy subset mapping, F-FKGs and retrieval code, partial implementation code for the detection module and MLLM explanation module, and training and evaluation scripts. It also provides frozen summaries of the detection and explanation metrics, together with a script that extracts the reported table values. Due to their size and supplementary upload limits, the current anonymous materials do not include raw images, model weights, feature caches, full prediction and judge outputs, or some internal model module files; complete end-to-end reproduction still requires these resources.

AI USE STATEMENT

We used generative AI to assist with code generation and editing, including implementation of the methods, to improve the manuscript's language and readability, and to generate or modify some scientific figures and images in the appendix. The authors reviewed all AI-assisted code and tested or otherwise validated it, reviewed all AI-edited manuscript content, and reviewed all AI-assisted appendix visuals to ensure that they accurately reflect the described methods, procedures, or experiments. The authors take full responsibility for the final text, code, experimental conclusions, and visual materials.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

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*, pp. 4113–4122, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Cao_End-to-End_Reconstruction-Classification_Learning_for_Face_Forgery_Detection_CVPR_2022_paper.html.

Yize Chen, Zhiyuan Yan, Guangliang Cheng, Kangran Zhao, Siwei Lyu, and Baoyuan Wu. $\mathcal{X}^2$-DFD: A framework for explainable and extendable Deepfake detection. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/78a2ed8e7da1829d64fb3561adbd706b-Abstract-Conference.html.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Chollet_Xception_Deep_Learning_CVPR_2017_paper.html.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

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*, pp. 9468–9478, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Dong_Protecting_Celebrities_From_DeepFake_With_Identity_Consistency_Transformer_CVPR_2022_paper.html.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A GraphRAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. doi: 10.48550/arXiv.2404.16130. URL https://arxiv.org/abs/2404.16130.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL https://jmlr.org/papers/v23/21-0998.html.

Xiang Gao and Kamalika Das. Customizing language model responses with contrastive in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18039–18046, 2024. doi: 10.1609/aaai.v38i16.29760. URL https://ojs.aaai.org/index.php/AAAI/article/view/29760.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 105–116, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Guo_Rethinking_Vision-Language_Model_in_Face_Forensics_Multi-Modal_Interpretable_Forged_Face_CVPR_2025_paper.html.

Hui Han, Shunli Wang, Yandan Zhao, Taiping Yao, and Shouhong Ding. VRAG-DFD: Verifiable retrieval-augmentation for MLLM-based Deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 9552–9562, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Han_VRAG-DFD_Verifiable_Retrieval-Augmentation_for_MLLM-based_Deepfake_Detection_CVPRF_2026_paper.html.

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. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zhengchao Huang, Bin Xia, Zicheng Lin, Zhun Mou, Wenming Yang, and Jiaya Jia. FFAA: Multimodal large language model based explainable open-world face forgery analysis assistant. *arXiv preprint arXiv:2408.10072*, 2024. URL https://arxiv.org/abs/2408.10072.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image Deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28831–28841, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Huang_SIDA_Social_Media_Image_Deepfake_Detection_Localization_and_Explanation_with_CVPR_2025_paper.html.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991. doi: 10.1162/neco.1991.3.1.79.

Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994. doi: 10.1162/neco.1994.6.2.181.

Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile GPUs. In *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, 2019. URL https://arxiv.org/abs/1907.06724.

Chenqi Kong, Anwei Luo, Peijun Bao, Yi Yu, Haoliang Li, Zengwei Zheng, Shiqi Wang, and Alex C. Kot. MoE-FFD: Mixture of experts for generalized and parameter-efficient face forgery detection. *IEEE Transactions on Dependable and Secure Computing*, 23(1):82–96, 2026. doi: 10.1109/TDSC.2025.3604443. URL https://doi.org/10.1109/TDSC.2025.3604443.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

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*, pp. 5001–5010, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Face_X-Ray_for_More_General_Face_Forgery_Detection_CVPR_2020_paper.html.

Tianxiao Li, Zhenglin Huang, Haiquan Wen, Yiwei He, Shuchang Lyu, Baoyuan Wu, and Guangliang Cheng. RAIDX: A retrieval-augmented generation and GRPO reinforcement learning framework for explainable Deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11746–11755, 2025. doi: 10.1145/3746027.3754798. URL https://doi.org/10.1145/3746027.3754798.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023. URL https://aclanthology.org/2023.emnlp-main.20/.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. URL https://arxiv.org/abs/1906.08172.

Aakash Varma Nadimpalli and Ajita Rattani. Facial forgery-based Deepfake detection using fine-grained features. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pp. 2174–2181, 2023. doi: 10.1109/ICMLA58977.2023.00328. URL https://arxiv.org/abs/2310.07028.

Tai D. Nguyen and Matthew C. Stamm. Trustworthy image authentication using forensic knowledge graphs. *arXiv preprint arXiv:2606.23917*, 2026. URL https://arxiv.org/abs/2606.23917.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ojha_Towards_Universal_Fake_Image_Detectors_That_Generalize_Across_Generative_Models_CVPR_2023_paper.html.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Computer Vision – ECCV 2020*, volume 12357 of *Lecture Notes in Computer Science*, pp. 86–103. Springer, 2020. doi: 10.1007/978-3-030-58610-2_6. URL https://link.springer.com/chapter/10.1007/978-3-030-58610-2_6.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Andreas Rössler, 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*, pp. 1–11, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.html.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Kaede Shiohara and Toshihiko Yamasaki. Detecting Deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18720–18729, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Shiohara_Detecting_Deepfakes_With_Self-Blended_Images_CVPR_2022_paper.html.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. Think-on-Graph: Deep and responsible reasoning of large language model on knowledge graph. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nnVO1PvbTv.

Hao Tan, Jun Lan, Zichang Tan, Senyuan Shi, Ajian Liu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. VERITAS: Generalizable Deepfake detection via pattern-aware reasoning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6d13e085b79d454da5910e4ca82a3d9d-Abstract-Conference.html.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2387–2395, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Thies_Face2Face_Real-Time_Face_CVPR_2016_paper.html.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/551d275ae5cb1bb1ebe6e37fbe201db8-Abstract-Conference.html.

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

International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=pAQzEY7M03.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. UCF: Uncovering common features for generalizable Deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22412–22423, 2023a. URL https://openaccess.thecvf.com/content/ICCV2023/html/Yan_UCF_Uncovering_Common_Features_for_Generalizable_Deepfake_Detection_ICCV_2023_paper.html.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. DeepfakeBench: A comprehensive benchmark of Deepfake detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 4534–4565, 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0e735e4b4f07de483cbe250130992726-Abstract-Datasets_and_Benchmarks.html.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70268–70288. PMLR, 2025. URL https://proceedings.mlr.press/v267/yan25b.html.

Peipeng Yu, Jianwei Fei, Hui Gao, Xuan Feng, Zhihua Xia, and Chip-Hong Chang. Unlocking the capabilities of large vision-language models for generalizable and explainable Deepfake detection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72925–72943. PMLR, 2025. URL https://proceedings.mlr.press/v267/yu25d.html.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhai_Sigmoid_Loss_for_Language_Image_Pre-Training_ICCV_2023_paper.html.

A HydraFake dataset

HydraFake (Tan et al., 2026) is a hierarchical out-of-distribution face-forgery dataset built to stress generalization beyond a single generator family. We use HydraFake's prescribed train/validation/test partition, following the same split as VERITAS rather than constructing a new one. Training examples come from this split; validation is used for model selection, calibration and threshold fitting, and test data are reserved for evaluation. The explanation SFT subset is drawn from its official split (§4.1 and Appendix E). Explanation evaluation uses 1,000 randomly sampled images. For a consistent detection benchmark, non-MEGRA results in Table 1 and the per-method detection tables are taken directly from VERITAS (Tan et al., 2026), reducing variation introduced by independently reimplementing or rerunning baselines on different software and hardware platforms. Our detection module is evaluated on the corresponding test subsets, subject to the FaceVid2Vid class-composition difference; the Score Evaluation results in Table 2 are produced in this study through a unified evaluation on 1,000 randomly sampled images. The test pool is partitioned into four protocols of increasing shift: ID (approximately 14k images), CM (11k), CF (12k), and CD (15k). ID covers classical Swap and Reenact (FF++, Hallo2, FaceVid2Vid) together with in-domain synthesis (Midjourney, StyleGAN). CM holds later generators that keep a related synthesis family but change the model (Adobe Firefly, FLUX, HART, Infinity, MAGI-1, StarryAI). CF mixes restoration, relighting, personalization and GAN editing (CodeFormer, IC-Light, FaceAdapter, PuLID, InfiniteYou, StarGANv2). CD includes commercial/social imagery and older in-the-wild sets (Dreamina, Hailuo, GPT-4o, DeepFaceLab, FFIW, InfiniteYou-CD). Real faces follow VERITAS (FF++ (Rössler et al., 2019) and later in-the-wild collections). Table 1 follows HydraFake's stratified reporting protocol in full. Method-level composition from VERITAS is summarized in Table 5.

Table 5: Subset composition of HydraFake's four test protocols, transcribed from VERITAS (Tan et al., 2026). FS=face swapping, FR=face reenactment, EFG=entire-face generation, AR=autoregressive.

Protocol	Method	Sub-type	Venue	Resolution
ID	FaceForensics++	FS	ICCV'19	256
	FaceVid2Vid	FR	arXiv'19	256
	Hallo2	FR	ICLR'25	256
	StyleGAN	EFG	CVPR'19	1024
	Midjourney	EFG	–	1024
CM	Adobe Firefly	Proprietary	–	1024
	StarryAI	Proprietary	–	1024
	Flux1.1 Pro	Customized	–	1024
	MAGI-1	Video AR	–	256
	HART	Image AR	arXiv'24	1024
	Infinity	Image AR	CVPR'25	1024
CF	StarGANv2	Editing	CVPR'20	256
	CodeFormer	Restoration	NeurIPS'22	512
	IC-Light	Relighting	ICLR'25	1536
	FaceAdapter	Generative FS	ECCV'24	1024
	PuLID	Personalization	NeurIPS'24	1024
	InfiniteYou	Personalization	ICCV'25	1024
CD	Hailuo AI	Commercial	–	256–1536
	Dreamina	Social media	–	1024
	GPT-4o	Social media	–	159–1536
	DeepFaceLab	Classic	PR'23	256
	FFIW	Classic	CVPR'21	256
	InfiniteYou-CD	Personalization	ICCV'25	1024

B EXPERIMENTAL SETUP

All experiments used one workstation; each run used one of four graphics processing units (GPUs), selected through `CUDA_VISIBLE_DEVICES`. No multi-GPU or distributed training was used. Tables 6 and 7 list the workstation configuration, SFT settings, and FD-MoE hyperparameters.

Table 6: Computing infrastructure (left) and explanation-model SFT hyperparameters (right). Abbreviations: central processing unit (CPU), random-access memory (RAM), operating system (OS), long-term support (LTS), learning rate (LR), and bfloat16 (bf16).

Component	Specification	Hyperparameter	Value
GPU	$4\times$ RTX 4090 (24 GB)	Base model	Qwen2.5-VL-7B (4-bit)
CPU	Xeon Gold 6430 (128 cores)	Adaptation	QLoRA r=32, α=64
System memory	251 GB RAM	Dropout	0.05
OS	Ubuntu 20.04.3 LTS	Epochs / LR	1 / 2.0×10^{-6}
Framework	PyTorch 2.6.0, CUDA 12.4	Batch / accum.	1 / 8
Mixed precision	bf16 autocast	Max seq. length	4096

Software environment. Python 3.10.20, PyTorch 2.6.0+`cu124`, torchvision 0.21.0+`cu124`, CUDA 12.4, the CUDA Deep Neural Network library (cuDNN) 9.1, Transformers 4.51.3, Accelerate 1.13.0, timm 1.0.27, NumPy 1.26.4, scikit-learn 1.7.2, Pillow 11.3.0, MediaPipe 0.10.21, PyYAML 6.0.3, and Parameter-Efficient Fine-Tuning (PEFT) 0.19.1.

Table 7: FD-MoE hyperparameters. The factorized head uses cross-entropy (CE). The decision threshold $\tau = -0.0039$ is fit on validation and frozen for the one-shot test pass.

Hyperparameter	Value	Hyperparameter	Value
Factorized-head hidden	256	Residual LoRA	r=16, α=32, fusion cap 0.04
Factorized-head loss	factorized CE	Residual epochs	8
Factorized-head opt. / seed	Adam 10^{-3} / 42	λ_{art}	0.8
Factorized-head epochs	20	α_{mix} / τ_{margin}	$[0.1, 0.6]$ / 0.05
Mechanism-expert MLP	1152→256→1, drop. 0.1	β / slack	1.0 / 0.05
Mechanism-expert ensemble	3 seeds	α_{synth} / α_{ID}	0.30 / 0.05
Expert-response hidden	32	δ_g / τ_g	0.12 / 0.02

The Refine constants in Table 7 instantiate Sec. 3.2.1: λ_{art} mixes p with π at family level; α_{mix} and τ_{margin} form a temperature-softmax of the margins m, with mix weight gated by the top-two margin gap through δ_g and τ_g; β and slack reallocate mass toward classes already on the fake side of τ (Appendix B.1). Residual scales α_{synth} and α_{ID} are those of Eq. (7). Bootstrap CIs of the frozen test predictions use 1,000 method-stratified resamples; extra factorized-head seeds are not part of the reported detection module. The photo-family residual uses α_{synth} and the synthesis LoRA score under the diffusion-class map, corresponding to $\mathbf{1}_{\text{photo}}$ in Eq. (7). Specifically, when the dominant mixture class is Diff/GAN and $F_c < \tau_c$, $\mathbf{1}_{\text{synth}} = 1$, and the residual adds the positive surpluses from both the identity-preserving and synthesis LoRA channels. When the dominant class is Swap/Reenact but $s_{\text{synth,Diff}} > F_c$, $\mathbf{1}_{\text{photo}} = 1$, and only the positive surplus from the synthesis LoRA is added; otherwise, $\Delta(x) = 0$.

B.1 CALIBRATION AND FACTORIZED-HEAD OBJECTIVES

The relationship between an expert's logit and authenticity can reverse between swaps and synthesis. We therefore calibrate each logit separately for each class as in Eq. (3), with polarity $a_{e,c} \in \{\pm 1\}$:

$$s_{e,c}(x) = Q_e\left(\sigma\left(\frac{a_{e,c} z_e(x) - b_{e,c}}{T_{e,c}}\right)\right), \tag{11}$$

where $T_{e,c} > 0$ and $b_{e,c}$ are class-specific, and the per-expert quantile map Q_e is shared across classes; all three are fit on validation. For class c, the polarity/temperature/bias pool is that class's fake validation images together with *all* real validation images. Polarity $a_{e,c}$ is set to -1 when the

AUC of the raw logit on that pool is below 0.5, and to $+1$ otherwise. The map Q_e is a monotone rank transform whose reference pool is the expert's calibrated validation scores over real and fake images jointly.

For the factorized head in Eq. (6), let $\mathcal{B}_{\mathrm{r}}$ be a minibatch of N fake training images with known mechanism labels c_i, and let $f_i = f(c_i)$ be their parent families. Swap and Reenact belong to photo-manipulation; diffusion and GAN synthesis belong to full synthesis. Labels are derived from training source methods, not test metadata; real images do not supervise $p(c \mid x)$. The head produces $p(c \mid x)$ and does not deactivate these channels: the frozen experts, mechanism experts, and SigLIP-LoRA channels all run, as in Sec. 3.2.1. The two cross-entropy terms are

$$\begin{aligned} \mathcal{L}_f &= -\frac{1}{N} \sum_{i \in \mathcal{B}_{\mathrm{r}}} \log p(f_i \mid x_i), \\ \mathcal{L}_{c|f} &= -\frac{1}{N} \sum_{i \in \mathcal{B}_{\mathrm{r}}} \log p(c_i \mid f_i, x_i). \end{aligned} \tag{12}$$

The first term supervises the two-way family head; the second supervises only the conditional head for the target family, with all samples averaged over the same minibatch. Under this factorization, the joint negative log-likelihood decomposes as $\mathcal{L}_f + \mathcal{L}_{c|f}$, so the objective is

$$\mathcal{L}_p = \mathcal{L}_f + \mathcal{L}_{c|f}. \tag{13}$$

Expert-response mixture distribution. The mixture distribution $\pi(c \mid x)$ is produced by a second factorized head with the same family/conditional factorization as Eq. (6) and hidden width 32. Its input is 10-dimensional: the six frozen-expert scores under class-agnostic calibration, concatenated with the four pre-offset margins $m_c + \delta_c$. Training uses Eq. (13) on fake training images; real images do not supervise π.

Mixture refinement. The family-level combination retains the within-family conditional probabilities of the visual factorized head p and incorporates the expert-response distribution π at the family level:

$$p'(c) \propto \Big[(1 - \lambda_{\mathrm{art}}) p(f(c) \mid x) + \lambda_{\mathrm{art}} \pi(f(c) \mid x)\Big] p(c \mid f(c), x). \tag{14}$$

For the two largest values $m_{(1)} \geq m_{(2)}$ of $m_c + \delta_c$, a margin-dependent coefficient interpolates within $[\alpha_{\mathrm{lo}}, \alpha_{\mathrm{hi}}]$:

$$\alpha = \alpha_{\mathrm{lo}} + (\alpha_{\mathrm{hi}} - \alpha_{\mathrm{lo}})\, \sigma\big(((m_{(1)} - m_{(2)}) - \delta_g)/\tau_g\big). \tag{15}$$

The family-level result is then combined with a softmax over the class margins:

$$p''(c) \propto (1 - \alpha)\, p'(c) + \alpha\, \mathrm{softmax}((m + \delta)/\tau_{\mathrm{margin}})_c. \tag{16}$$

Using the threshold τ, a fraction β of mass on classes with $m_c < \tau$ is then moved onto synthesis classes whose margins already exceed $\tau + \mathrm{slack}$: from Swap onto $\{\mathrm{Diff}, \mathrm{GAN}\}$ in proportion to their surplus when at least one synthesis margin is above the slack line, and from GAN onto Diff when m_{Diff} is above the slack line while $m_{\mathrm{GAN}} < \tau$. Norm renormalizes the result onto the simplex, yielding $\tilde{p}$.

Class-conditional fusion. Sparse nonnegative weights $w_{c,h}$ are fit by coordinate ascent on leave-one-method-out validation BA. Each coordinate is searched over the 21-point grid $\{0, 1/20, \ldots, 1\}$; a round updates every channel in turn, and search stops after at most 8 rounds or when a round does not improve the objective. After every trial the weights are projected onto the simplex with the following constraints: an Xception floor of 0.15; mechanism-expert caps of 0.22 (Swap), 0.55 (Reenact), 0.55 (diffusion), and 0.40 (GAN); a per-channel cap of 0.40 under GAN; Swap floors of 0.203 on Effort and 0.08 on UCF; and a synthesis-LoRA fusion cap of 0.04 on diffusion and GAN. A channel receives weight zero when the search selects 0; there is no separate utility cutoff. Each class threshold τ_c is the BA-optimal cut on that class's fused validation scores after the weights are fit. The Swap offset $\delta_{\mathrm{Swap}} = 0.03$ (other $\delta_c = 0$) is chosen on validation to control false positives on that head and is frozen at test time.

B.2 DETECTOR ABLATION DEFINITIONS

Table 3 uses cached expert logits and fixed detection-module checkpoints. Each module lesion zeros or disables the named block, recomputes intermediate scores together with p, π, and $\tilde{p}$, and refits the decision threshold on validation by BA; networks, polarity/temperature/bias maps, quantile maps, and fusion weights are not retrained. Removing fusion channels zeros their weights and renormalizes the remaining nonnegative weights within each class. ID/CM/CF/CD in the table are unweighted means of the corresponding method groups, matching Table 3.

Frozen experts. *w/o frozen experts* zeros the six forensic channels in every class-conditional fusion vector and turns off the expert-response mixture distribution π, which is defined on those responses. Mechanism experts, the factorized head p, and residual LoRA remain.

Mechanism experts. *w/o mechanism experts* removes all four mechanism-expert heads and their fusion weights, while retaining frozen SigLIP features, both factorized-head branches (family and within-family), and residual LoRA channels.

Class calibration. *w/o class calib.* resets polarity, temperature, and bias to $a_{e,c} = 1$, $T_{e,c} = 1$, and $b_{e,c} = 0$ for every calibration entry, including class-agnostic entries used to form π. The sigmoid and existing quantile maps are retained, giving $s_{e,c} = Q_e(\sigma(z_e))$ wherever quantile normalization is enabled. Fusion weights remain fixed; the threshold is then refit.

Max-margin. This variant replaces the mixture-weighted margin with the largest class margin and keeps the residual. The threshold is refit on the max-margin scores.

Residual. The residual row disables the SigLIP-LoRA detection score channels: it zeros their class-conditional fusion weights and turns off $\mathbf{1}_{\text{synth}}$ and $\mathbf{1}_{\text{photo}}$ in Eq. (7). This jointly removes the LoRA fusion weights and Δ.

Frozen-pool baselines. Single experts and both aggregation baselines in Table 17 use each expert's class-agnostic Platt-calibrated probability, fitted on validation without mechanism-conditioned polarity, temperature, or bias. *Uniform avg. of 6* is the arithmetic mean of these six probabilities. *Global stack of 6* fits one logistic regressor with an intercept to the same six-dimensional vector, using class-balanced binary cross-entropy with an ℓ_2 penalty on validation. Its weights are shared across all mechanisms. Each baseline selects its threshold by validation BA; test labels are used only for reporting. These rows are frozen expert-pool baselines using released weights; the standalone detectors in Table 1 follow their respective HydraFake training settings (Appendix J).

C FROZEN EXPERTS AND REGIONAL EVIDENCE

The six frozen experts are independently trained forensic systems; we do not fine-tune them. Table 8 lists the paper, backbone, parameter count (from the deployed checkpoints, excluding batch-norm running statistics) and the forensic cue each expert is intended to capture. Xception, UCF and RECCE use DeepfakeBench weights (Yan et al., 2023b); SBI, UnivFD and Effort use the original releases. The six expert networks remain frozen during detection-module training (Table 7). A DeepfakeBench EfficientNet-B4 (17.6M) and F3Net were evaluated but received zero fusion weight and were omitted from the official pool.

Regional evidence. We use MediaPipe (Lugaresi et al., 2019) Face Mesh (Kartynnik et al., 2019) for facial alignment. For each of eight regions (eyes, nose, mouth, skin, jawline, `face_boundary`, hairline, background), we pool a 512-d descriptor and a class-activation energy from the frozen Xception sidecar's spatial map. A geometric face-crop fallback is used if landmark detection fails. Region scores form R_x and seed GraphRAG (Appendix D); they are not fused into $S(x)$. Only the Xception sidecar provides these regional features; the other five experts each emit a scalar logit into the mixture.

Table 8: Sources, backbones, cues, and parameter counts of the six frozen forensic experts. Parameter counts include frozen parameters in the deployed checkpoints. Abbreviations: contrastive language–image pre-training (CLIP) and singular value decomposition (SVD). For our deployed UnivFD checkpoint, the CLIP ViT-L/14 backbone (Radford et al., 2021) is frozen and the trainable linear classifier contains 769 parameters; the reported 304M includes the frozen vision encoder. Xception, UCF and RECCE are DeepfakeBench checkpoints (Yan et al., 2023b).

Expert	Paper	Backbone	Cue	#Params
Xception	CVPR'17 (Chollet, 2017)	Xception CNN	Generic spatial CNN	21.9M
UCF	ICCV'23 (Yan et al., 2023a)	Dual Xception encoders	Common vs. specific	46.8M
RECCE	CVPR'22 (Cao et al., 2022)	Xception encoder–decoder	Reconstruction residual	47.7M
SBI	CVPR'22 (Shiohara & Yamasaki, 2022)	EfficientNet-B4	Blend-boundary traces	17.6M
UnivFD	CVPR'23 (Ojha et al., 2023)	CLIP ViT-L/14 + linear	Foundation features	304M
Effort	ICML'25 (Yan et al., 2025)	CLIP ViT-L/14 + SVD residual	Orthogonal subspace	303M

D F-FKGs AND GRAPHRAG

Figure 4 summarizes the typed F-FKGs used by region-guided GraphRAG. Here GraphRAG denotes retrieval over graph-structured forensic knowledge to augment explanation generation. The diagram is schematic: it highlights the main node layers, evidence paths, and how retrieved paths condition the SFT MLLM.

The staged flow (region → cue → forensic rule / authenticity pattern → category or real pattern → supporting text) summarizes the main conceptual hierarchy; the retrieval units are defined below. The three headline forgery categories (face swap, reenactment, entire-face generation) and the real authenticity branch align with the supervised explanation targets in Sec. 3.2.2. The compiled graph additionally contains eight finer typed families. The figure shows a representative subset of seed regions and cues; the complete graph contains 21 region nodes and 27 cue nodes. The green "authentic" entries in the cue column correspond to authenticity-pattern nodes. The two example explanations in the rightmost column are produced by the SFT-trained Qwen2.5-VL-7B explanation module.

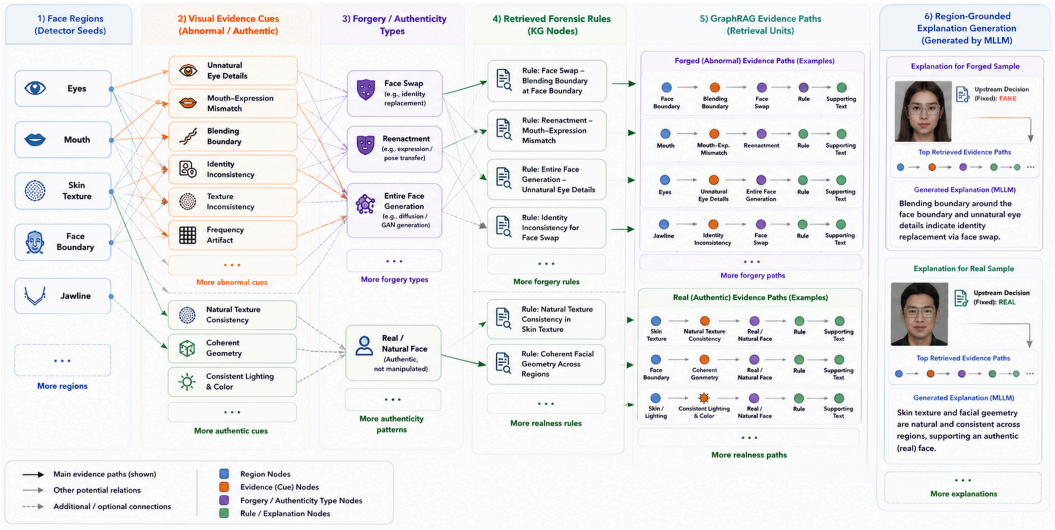


Figure 4: Schematic of F-FKGs and region-guided GraphRAG retrieval. Columns 1–4 give conceptual layers: facial regions (blue), candidate cues and authenticity patterns (orange/green), forgery categories and the real-face branch (purple), and forensic rules with supporting text. Column 5 illustrates evidence paths in conceptual order; real retrieval units contain region, authenticity-pattern, and text nodes. Column 6 shows decision-conditioned explanation. Solid arrows highlight selected evidence; thin and dashed arrows show additional conceptual associations.

D.1 F-FKGs INVENTORY

Table 9 lists the compiled F-FKGs statistics. All retrievable nodes exclude generator names, source methods, file paths, dataset splits, sample identifiers, and per-sample ground-truth labels.

Table 9: Compiled F-FKGs inventory (260 nodes in total).

Node type	Role	Count
facial region	Canonical face / scene region anchors	21
visual cue	Candidate abnormal or ambiguity cues	27
evidence rule	Reusable region–cue forensic rules (fake)	88
forgery category	Typed manipulation families	11
authenticity pattern	Positive-evidence patterns (real)	3
caution policy	Conservative-support policies	4
text chunk	Supporting forensic text for rules/patterns	106
Edges (typed relations)		611

D.2 REGIONS, CUES, AND RELATIONS

Detection-module seed regions. The frozen Xception sidecar exports eight MediaPipe-aligned seed regions that anchor retrieval (Appendix C): eyes, nose, mouth, skin, jawline, `face_boundary`, hairline, and background. These map to 21 canonical facial-region nodes, including finer structures (e.g., lips, teeth, cheeks, forehead, lighting) that rules may reference without changing the detection module's interface. Region scores are not fused into $S(x)$.

Visual cues and cue groups. Retrieval uses 27 canonical visual-cue nodes grouped into eight cue groups: boundary blending, texture/color, lighting/shadow, geometry/structure, identity/expression/pose, generation/frequency, background/context, and quality/ambiguity. Candidate cues are generated deterministically from high-response regions through a fixed region-to-cue map and used to condition GraphRAG queries.

Edge types. Relations are typed as localized-at, associated-with, supports, described-by, requires-caution, contradicts, and co-occurs-with. Fake rules connect a region and cue to a forgery category and a supporting text chunk; authenticity patterns connect regions to real-support text; caution policies connect ambiguity cues to conservative guidance.

D.3 EVIDENCE PATH TEMPLATES

Fake paths. For a fake decision, a retrieved path has the form of Eq. (8), where r is a facial region, u a candidate cue, η an evidence rule, t a compatible forgery category, and d supporting text. The rule η is the hub of all four edges, so the evidence path forms a typed star subgraph; the reading order in the figure presents its conceptual hierarchy.

Real paths. Real-image retrieval connects a seed region to an authenticity pattern and supporting text, e.g. global consistency, texture–geometry coherence, or lighting–context coherence. A real path comprises a region, an authenticity pattern, and supporting text; the green "authentic" nodes in the figure denote these pattern-level nodes.

Caution paths. When confidence is limited or ambiguity cues such as compression, low resolution, or blur are present, caution paths attach a corresponding policy and supporting text so that generation remains consistent with the strength of the evidence.

D.4 RETRIEVAL, RANKING, AND PROVENANCE

Retrieval is conditioned on the detection module's fixed decision (Eq. (2)), high-response regions, cue hypotheses, and confidence state. A coarse type hint from $\arg\max_c \tilde{p}_c$ conditions retrieval as a detection-side prior. A cue proposal map Ψ forms inspection priors $C_x = \Psi(\mathrm{Top}(R_x), \hat{y})$ from the highest-energy regions and the decision, matching Sec. 3.2.2. Candidate fake paths must overlap

regional seeds and either an exact cue or its canonical cue group. Ranking uses Eq. (9) with the seven features $q = (D_p, O_r, O_c, I_f, I_h, U_p, S_p)$ and $\lambda = (2.0, 1.5, 1.5, 1.0, 0.8, 0.8, 0.5)$. Forbidden generator, split, and sample-identifier fields are removed when the graph is built, so they are not a ranking coordinate. Decision compatibility D_p is 1 for a match, 0.35 for an uncertainty-compatible path, and 0 otherwise. Region overlap O_r is the fraction of query regions covered; cue overlap O_c takes the maximum of exact-cue and cue-group coverage. Rule/text features I_f and I_h are presence indicators. The uncertainty feature U_p is 1 for a caution path under medium/high query uncertainty and 0.2 otherwise. Evidence strength S_p is 0.2/0.6/1.0 for weak/medium/strong. The retriever keeps at most five paths with at most four hops, and limits repeated regions and cue groups to two before fallback filling. Selected path identifiers are recorded in the explanation provenance after generation.

E SFT DETAILS

Conditional generation. For example i, write $\mathcal{I}_i = (x_i, \hat{y}_i, R_{x_i}, \mathcal{P}_{x_i})$ for the conditioning input: the image, fixed detection-module decision, regional evidence, and retrieved knowledge paths; the instruction and output schema are implicit. The supervised target $Y_i = (u_{i,1}, \ldots, u_{i,T_i})$ is the serialized assistant response, including its structured fields and explanation text. Its autoregressive likelihood is

$$\pi_\theta(Y_i \mid \mathcal{I}_i) = \prod_{t=1}^{T_i} \pi_\theta(u_{i,t} \mid \mathcal{I}_i, u_{i,<t}). \tag{17}$$

Training uses teacher forcing, conditioning on the reference prefix $u_{i,<t}$; inference instead conditions on previously generated tokens.

Masked, token-weighted objective. The implemented masked, token-weighted loss is given in Eq. (10). Here $m_{i,t}$ is one for a valid assistant-target token and zero for padding; prompt tokens are excluded from supervision before next-token alignment. The denominator is the total valid-token weight in the minibatch, with a numerical floor of 10^{-6}; batches without supervised tokens are rejected. Weights $w_{i,t}$ are preset: a token-level heuristic assigns 0.3 to nonempty tokens containing only JSON formatting characters after whitespace stripping, 0.5–0.7 to complete key-like tokens, and 1.0 otherwise. The weights are computed at the decoded-token level; ordinary explanation-content tokens retain weight 1.0. Setting all valid-token weights to one recovers token-average negative log-likelihood.

Residual evidence injection. Each of the eight Xception regions contributes a 512-d pooled descriptor concatenated with its class-activation energy (513-d). A global descriptor concatenates a 512-d pooled map, the sidecar fake probability, and a 4-d detection-module prior (real score, fake score, margin, absolute margin), giving a 517-d vector. Two residual MLPs, each $\mathrm{LayerNorm} \to \mathrm{Linear} \to \mathrm{GELU} \to \mathrm{Linear}$, map these inputs to the language-model hidden size (3584 for Qwen2.5-VL-7B). Learned scalars α and γ scale the global and region residuals, which are added to the embeddings of special tokens placed in the text prompt: one global token $\langle \mathrm{EVIDENCE_SUMMARY} \rangle$ and at most the three highest-energy region tokens, rather than one token per region. Image tokens, the text prompt, and retrieved GraphRAG path text remain ordinary multimodal inputs; only these evidence tokens receive the residual.

Trainable parameters and decision preservation. The SFT implementation updates the language-model QLoRA adapters, the residual evidence-injection module, and the multimodal projector; the remaining pretrained backbone parameters and upstream detection module are frozen. Equation (2) preserves the upstream decision as an output constraint and introduces no additional differentiable loss. At inference, the generated response is checked against this contract; if the check fails, the response is regenerated once with a repair instruction (Appendix K), after which the delivered decision field is set to $\hat{y}_i$.

Data and settings. MLLM SFT uses 32,870 training rows and 1,612 validation rows; 5,366 teacher rewrites are counted across both splits. Infrastructure and QLoRA settings are in Table 6; the QLoRA targets are $q/k/v/o/\mathrm{gate}/\mathrm{up}/\mathrm{down}$. Remaining SFT settings (maximum pixels and sequence length) match the frozen explanation checkpoint. Input ablations use the same SFT check-

point without retraining: removing F-FKGs clears retrieved context; additional region removal clears region/cue candidates and zeros the regional evidence tokens, including their global summary.

F TRAINING AND INFERENCE COST

We compare wall-clock training time, single-decode inference latency, trainable-parameter counts and peak GPU memory with VERITAS (Tan et al., 2026). Detection accuracy is not re-measured here; Table 1 already reports the HydraFake numbers.

Training time. VERITAS reports a full-pipeline training time of 18.5 hours on eight PPUE GPUs (96 GB each): 5.5 hours of SFT, 2 hours of MiPO, and 11 hours of P-GRPO. MEGRA's full-pipeline training time on a single RTX 4090 (24 GB) is 13.4 hours, measured from job launch to the official checkpoint. The detection module—a factorized head, four mechanism-expert MLPs and two residual SigLIP-LoRA channels—is trained on the same single GPU from cached features and does not use a multi-GPU MLLM schedule.

Inference latency. All timings in Table 10 use one RTX 4090, batch size 1 and the same 24 validation images. MEGRA can serve the detection module without starting the explanation module. The measured fusion-to-decision stage takes 3.9 ms after frozen-expert scores and SigLIP features are available; this timing excludes their forward passes and image preprocessing. VERITAS jointly emits the verdict and the rationale in one `chat` (9.24 ± 2.57 s, bf16) and has no detection-only serving path. When compute allows, the explanation module can be enabled: a single greedy decode takes 11.5 s (4-bit NF4; `max_new_tokens=768`; no JSON-contract retry). The modest gap relative to VERITAS is due to 4-bit NF4 decoding, the longer GraphRAG and JSON-contract prompt, and residual evidence injection through `inputs_embeds`. Official serving with one JSON-contract retry is slower (23.8 s) and is not used in this comparison.

Parameters and memory. VERITAS does not print a trainable-parameter integer. Instantiating its LoRA recipe on the released InternVL3-8B graph with `freeze_vit=false` (the authors' public SFT script) yields 373.3M trainable adapters; the merged inference checkpoint contains 7.94B parameters. MEGRA trains 137.1M parameters in total (12.5M on the detection module and 124.6M on the explanation module), about one third of that LoRA count, while remaining competitive on HydraFake (Table 1). Frozen forensic experts, including CNN and ViT backbones (Table 8), and the frozen SigLIP encoder are not counted as trainable; a detection-only deployment also loads these backbones alongside the 12.5M trainable detector components, without the 7B explanation module. Peak explanation-generation memory is 6.73 GB (4-bit) versus 18.4 GB for VERITAS (bf16) under the measured decode settings.

Table 10: Training resources, parameter counts, and inference costs of MEGRA and VERITAS. VERITAS training hours are transcribed from Tan et al. (2026). Latency and peak memory are measured on one RTX 4090 over the same 24 validation images with batch size 1; the fusion-to-decision and joint-decode latency rows cover different stages. †Trainable LoRA count instantiated from the VERITAS recipe; not a value printed in the VERITAS paper.

	MEGRA	VERITAS
Training GPUs	1× RTX 4090 (24 GB)	8× PPUE (96 GB)
Training wall-clock	13.4 h (full pipeline)	18.5 h (full pipeline)
Trainable parameters	137.1M	373.3M†
Fusion-to-decision latency	3.9 ms	— (joint decode only)
Explanation / joint, single decode	11.5 s (4-bit)	9.24 s (bf16)
Peak generation GPU memory	6.73 GB (4-bit)	18.4 GB (bf16)

G HYPERPARAMETER SENSITIVITY

We probe six Refine and residual scalars from Table 7 one at a time on the frozen official detection module, keeping the validation threshold τ fixed. Defaults were selected on validation before this

sweep; test numbers here are not used to change the official configuration. Equation (7) uses a shared α_{synth}; the sweep varies that scale separately on the $\mathbf{1}_{\text{synth}}$ and $\mathbf{1}_{\text{photo}}$ branches. The upper endpoint of $\alpha_{\text{mix}} = [0.1, 0.6]$ is the scalar that is swept.

Table 11 reports the range of average accuracy (the same Avg. as Table 1, not the mean of four protocol sets). The default remains 91.09% throughout. Except when $\alpha_{\text{ID}} = 0.12$, varying four scalars—$\lambda_{\text{art}}$, the upper end of α_{mix}, β, and α_{ID}—changes average accuracy by at most 0.15 points from the default. Setting the synthesis-branch α_{synth} to 0 drops the mean to 90.32; setting $\alpha_{\text{ID}} = 0.12$ drops it to 90.35, mainly on ID and CM. A synthesis-branch value 0.4 reaches 91.13 on this frozen test pass; we do not adopt it.

Table 11: One-at-a-time sensitivity of Refine and residual scalars on the frozen detection module (average accuracy, %). τ is held at the official validation value. †Separate branch of the shared α_{synth} in Eq. (7).

Scalar	Default	Values swept	Min–max
λ_{art}	0.8	$0, 0.4, 0.6, 0.8, 1.0$	90.99–91.09
α_{mix} (upper end)	0.6	$0, 0.3, 0.6, 0.8, 1.0$	90.94–91.09
β	1.0	$0, 0.5, 1.0$	90.94–91.09
α_{ID}	0.05	$0, 0.02, 0.05, 0.08, 0.12$	90.35–91.09
α_{synth} $(\mathbf{1}_{\text{synth}})^{\dagger}$	0.30	$0, 0.1, 0.2, 0.3, 0.4$	90.32–91.13
α_{synth} $(\mathbf{1}_{\text{photo}})^{\dagger}$	0.30	$0, 0.1, 0.2, 0.3, 0.4$	90.69–91.10

H PAIRED STATISTICAL SIGNIFICANCE

A paired statistical significance test compares two predictors on the *same* images: each sample contributes one pair of outcomes. Each comparison pairs the official MEGRA detection module with one module lesion from Table 3, aligned on all $53{,}155$ HydraFake test images.

Table 3 reports average accuracy, the same Avg. as Table 1, not the mean of the four protocol columns. McNemar's test, by contrast, acts on per-image Real/Fake decisions (micro-accuracy). We therefore use two complementary paired procedures while keeping the full model's official operating point fixed. Lesion decisions reuse the validation-refit threshold already used in Table 3.

McNemar on decisions. Let n_{10} be the number of images that MEGRA classifies correctly and the lesion incorrectly, and n_{01} the converse. The test asks whether n_{10} and n_{01} are exchangeable. We use the continuity-corrected χ^2 form and Holm-adjust the five lesion p-values.

Paired bootstrap of the average-accuracy gap. We resample images within each source method, recompute both average accuracies on that draw, and form Δ = MEGRA minus lesion (percentage points). Table 12 reports the percentile 95% interval of Δ from $1{,}000$ replicates. This interval measures the average-accuracy gap between the full model and each variant.

Every drop in Table 12 is significant: all intervals exclude 0, and every Holm-adjusted McNemar p is below 0.001. Removing mechanism experts yields the largest gap (6.05 points); Max-margin yields the smallest (1.16 points) while remaining significant. These tests concern the internal detection-module lesions only.

I ADDITIONAL METRICS

MEGRA's per-image (micro) accuracy (%) / AUC values are ID 93.21 / 0.976, CM 97.49 / 0.994, CF 89.51 / 0.957, and CD 79.42 / 0.855, distinct from the method-macro values in Table 1. The bootstrap 95% CI for average accuracy is $[90.77, 91.42]$ (1,000 method-stratified resamples). The detection module's factorized head and four mechanism-expert MLPs contain 3.84M trainable parameters, excluding separately trained residual SigLIP-LoRA adapters and the explanation-side QLoRA, evidence-injection, and projector parameters. A CPU microbenchmark measures mixture/fusion latency at 0.38 ms/image, excluding frozen expert forward passes. On one RTX 4090, a single expla-

Table 12: Paired tests of MEGRA against the lesions in Table 3. Avg. is the lesion's average accuracy (%). Δ is MEGRA minus lesion (percentage points). The 95% CI is for Δ. n_{10}/n_{01}: MEGRA correct and lesion wrong / the converse. p: Holm-adjusted McNemar.

Lesion	Avg.	Δ	95% CI of Δ	n_{10}/n_{01}	p
w/o mechanism experts	85.04	6.05	$[5.63, 6.47]$	$4320/1564$	< 0.001
w/o frozen experts	87.51	3.58	$[3.20, 3.94]$	$6109/2019$	< 0.001
w/o class calib.	88.84	2.25	$[1.94, 2.55]$	$2321/1179$	< 0.001
w/o residual	88.91	2.18	$[1.92, 2.45]$	$1241/488$	< 0.001
Max-margin	89.93	1.16	$[0.87, 1.41]$	$1530/877$	< 0.001

nation decode takes 11.5 s (Appendix F). Detector component and expert-count ablations matching Table 3 are included in the anonymous supplementary code.

I.1 Per-method Acc., Precision, and Recall

Let fake be the positive class. We denote true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN); N is the number of images. We report accuracy (Acc.), precision (P.), and recall (R.): $\mathrm{Acc} = (\mathrm{TP}+\mathrm{TN})/N$, $\mathrm{Precision} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, and $\mathrm{Recall} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$. A score greater than or equal to the frozen validation threshold is classified as fake. Tables 13–16 give per-method Acc./P./R. under ID/CM/CF/CD. Except for the MEGRA row, every entry is transcribed from VERITAS (Tan et al., 2026); the last row is our method on the full test split (same operating point as Table 1). Avg. is the unweighted mean of per-method Acc. **Bold**/underline: best/second-best per column, including ties.

Table 13: ID Acc./P./R. on HydraFake. Except the MEGRA row, all entries are transcribed from VERITAS (Tan et al., 2026). The last row is MEGRA on the full ID test set (same global validation threshold as Table 1). Columns report Accuracy, Precision, and Recall for FaceForensics++, FaceVid2Vid, Hallo2, StyleGAN, and Midjourney. Avg. is the unweighted mean of the five per-method Acc. values (VERITAS protocol), which can differ slightly from the sample-weighted ID aggregate above. MEGRA's FaceVid2Vid subset contains 1,000 fake images and no real images; its Acc. and P. are therefore not directly comparable with the published baselines on that subset. The AIDE StyleGAN Acc./P./R. values are reproduced as reported in VERITAS, although they do not form a consistent single confusion matrix. **Bold**/underline: best/second-best per column, including ties.

Method	FF++			FaceVid2Vid			Hallo2			StyleGAN			Midjourney			Avg.
	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	
F3Net	84.5	80.5	90.9	89.6	83.6	98.6	85.0	78.3	97.0	82.5	77.4	91.6	84.8	78.6	95.6	85.3
UnivFD	73.8	80.4	62.8	81.7	83.2	79.6	77.3	88.8	62.5	94.3	91.5	97.6	86.3	90.6	81.0	82.7
IID	83.8	86.6	79.9	92.8	89.5	97.0	79.1	72.2	94.6	80.7	72.1	**100.0**	80.5	71.9	**100.0**	83.4
FreqNet	52.2	51.2	94.4	54.5	52.3	**100.0**	71.2	66.7	84.5	77.5	69.8	96.6	78.8	70.8	98.0	66.8
ProDet	84.8	82.2	88.8	90.9	84.6	**100.0**	91.4	89.0	94.4	92.5	88.7	97.3	93.0	88.4	99.0	90.5
NPR	59.6	56.5	84.1	66.6	60.5	95.7	74.3	87.2	57.0	88.8	91.7	85.3	88.7	92.9	83.6	75.6
AIDE	58.9	58.7	59.6	78.6	70.3	99.0	84.5	92.4	75.2	86.8	93.1	99.0	93.0	92.7	93.3	80.4
Co-SPY	71.3	76.0	62.3	89.3	82.9	99.1	82.1	91.6	70.7	95.0	92.2	98.3	94.0	94.0	94.0	86.3
D^3	72.5	70.3	77.8	82.1	74.7	96.9	91.1	91.7	90.3	96.3	94.2	98.6	94.7	91.3	98.6	87.3
Effort	**92.9**	**94.4**	91.2	96.4	94.6	98.4	95.9	95.3	96.5	95.0	97.5	92.3	93.5	96.1	90.6	94.7
Qwen2.5-VL-7B	51.3	93.5	2.8	51.5	94.3	3.3	50.0	50.0	1.1	51.5	80.0	4.0	51.5	76.4	4.3	51.2
InternVL3-8B	55.9	55.5	59.0	53.7	53.6	54.6	56.3	83.8	15.6	52.0	66.6	8.0	52.3	69.4	8.3	54.0
MiMo-VL-7B	56.2	55.3	64.0	62.2	59.4	77.1	60.4	63.9	47.6	66.4	69.3	58.6	73.6	72.0	77.3	63.8
GLM-4.1V-T	58.0	64.0	36.3	59.6	65.6	40.0	54.2	89.6	9.4	56.0	87.5	14.0	54.3	80.9	11.3	56.4
GPT-4o	49.2	20.0	0.5	53.0	84.2	8.0	49.8	33.3	0.5	63.2	**100.0**	26.5	52.5	91.6	5.5	53.5
Gemini-2.5-Pro	62.0	77.9	33.5	53.2	66.6	13.0	66.0	80.9	42.5	93.9	89.3	**100.0**	85.8	73.7	76.6	72.2
VERITAS	90.9	90.2	91.7	96.1	92.7	**100.0**	**99.8**	**99.6**	**100.0**	**99.8**	99.6	**100.0**	**100.0**	**100.0**	**100.0**	**97.3**
MEGRA	92.2	89.0	**94.6**	**97.3**	**100.0**	97.3	93.6	96.9	90.0	97.2	94.6	**100.0**	97.3	94.9	**100.0**	95.5

Table 14: CM Acc./P./R. on HydraFake. Except the MEGRA row, all entries are transcribed from VERITAS (Tan et al., 2026) (ADF, FLUX, StarryAI, MAGI-1, HART, Infinity). The last row is MEGRA on the full CM test set (same global validation threshold as Table 1). Avg. is the unweighted mean of the six unrounded per-method Acc. scores. **Bold**/underline: best/second-best per column, including ties.

Method	ADF			FLUX			StarryAI			MAGI-1			HART			Infinity			Avg.
	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	
F3Net	86.7	80.0	97.7	87.8	80.4	**100.0**	78.6	76.2	83.3	85.0	78.6	96.2	86.0	78.6	99.0	82.9	78.6	90.4	84.5
UnivFD	90.7	92.9	88.0	93.8	91.7	96.3	82.5	88.8	74.3	73.0	84.7	56.1	94.4	92.2	97.0	90.7	92.8	88.2	87.5
IID	83.3	75.0	**100.0**	82.8	74.4	**100.0**	80.0	72.1	**97.6**	80.2	72.6	96.5	81.1	72.6	**100.0**	82.2	73.7	**100.0**	81.6
FreqNet	60.3	59.5	64.3	76.7	68.4	99.0	59.0	59.1	58.3	69.2	65.5	81.1	77.1	69.4	96.7	75.1	68.7	92.3	69.6
ProDet	92.6	88.3	98.3	94.2	89.5	**100.0**	88.2	88.0	88.3	91.9	87.5	97.7	93.8	89.0	99.8	93.1	88.7	98.6	92.3
NPR	68.8	86.0	45.0	91.2	91.5	90.6	59.5	78.8	26.0	82.6	91.9	71.5	91.3	93.1	89.3	84.0	92.6	73.8	79.6
AIDE	68.8	85.5	43.3	86.3	91.0	80.6	64.0	86.6	33.0	88.9	93.6	83.6	95.4	94.3	96.6	76.0	91.0	57.7	79.9
Co-SPY	93.5	93.6	93.3	95.5	94.1	97.0	85.3	93.4	76.0	93.3	92.2	72.7	96.6	94.5	98.9	95.3	94.1	96.7	93.3
D^3	93.6	90.3	97.7	95.6	92.3	99.6	91.3	92.2	90.3	90.7	90.5	91.0	95.8	92.2	99.9	95.5	93.1	98.3	93.8
Effort	82.8	97.1	67.6	96.5	96.0	97.0	78.0	94.2	59.6	90.5	96.5	84.1	97.8	97.2	98.5	98.3	97.3	99.4	90.7
Qwen2.5-VL-7B	50.0	50.0	1.3	50.0	50.0	1.3	49.7	0.0	0.0	50.0	50.0	0.9	52.0	83.8	4.9	52.9	87.3	6.8	50.8
InternVL3-8B	54.0	74.0	12.3	49.8	47.3	3.0	49.0	28.5	1.3	56.6	83.5	16.4	55.8	83.1	14.5	57.2	84.5	17.6	53.7
MiMo-VL-7B	74.5	74.7	74.0	77.1	76.7	78.0	82.5	68.0	55.3	60.3	64.0	47.1	82.4	77.9	90.6	81.4	77.2	89.2	76.4
GLM-4.1V-T	55.2	82.9	13.0	52.3	76.9	6.6	50.5	66.6	2.0	51.6	81.5	4.2	68.4	96.1	38.2	60.7	95.5	22.5	56.5
GPT-4o	57.7	96.9	16.0	52.0	83.3	5.0	51.4	85.7	3.0	59.9	80.0	26.4	81.2	92.5	67.9	54.8	87.5	10.6	59.5
Gemini-2.5-Pro	64.9	78.8	41.2	92.4	90.3	94.9	82.8	92.3	72.0	62.5	79.0	34.1	93.4	88.5	**100.0**	93.2	89.1	98.5	81.5
VERITAS	94.8	**99.6**	89.9	**99.8**	**99.6**	**100.0**	**97.0**	**100.0**	94.0	**99.9**	**99.8**	**100.0**	**99.9**	**99.8**	**100.0**	**99.9**	**99.8**	99.9	**98.6**
MEGRA	**97.5**	97.0	98.0	97.3	95.8	99.0	91.5	93.7	89.0	98.1	96.5	99.8	97.7	95.7	**100.0**	98.0	96.4	99.6	96.7

Table 15: CF Acc./P./R. on HydraFake. Except the MEGRA row, all entries are transcribed from VERITAS (Tan et al., 2026) (StarGANv2, IC-Light, CodeFormer, InfiniteYou, PuLID, FaceAdapter). The last row is MEGRA on the full CF test set (same global validation threshold as Table 1). Avg. is the unweighted mean of the six per-method Acc. scores. **Bold**/underline: best/second-best per column, including ties.

Method	StarGANv2			IC-Light			CodeFormer			InfiniteYou			PuLID			FaceAdapter			Avg.
	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	
F3Net	41.3	20.7	6.2	48.9	47.9	24.8	71.9	72.8	69.8	84.9	78.2	96.8	85.5	79.5	95.7	72.6	77.5	62.8	67.5
UnivFD	61.8	81.7	30.4	81.9	89.1	72.6	75.4	88.0	58.7	73.7	87.0	55.6	68.1	83.6	45.1	81.3	90.3	69.6	73.7
IID	41.4	32.8	16.6	53.3	54.3	41.0	79.7	72.9	94.4	81.8	73.4	**99.6**	81.8	73.3	**99.9**	73.7	68.2	87.1	68.6
FreqNet	33.1	11.4	5.0	73.1	67.6	88.9	70.3	66.6	81.5	72.8	67.4	88.3	77.4	69.9	96.2	67.7	64.9	75.0	65.7
ProDet	56.3	67.5	24.5	58.6	69.8	34.4	80.8	84.7	75.3	88.1	89.9	85.8	91.0	88.1	94.9	83.3	86.0	79.0	76.4
NPR	47.7	23.6	2.1	67.8	85.2	43.2	60.6	78.3	29.3	79.8	99.1	66.0	89.0	92.8	84.5	67.7	84.9	41.9	68.8
AIDE	56.7	74.6	20.3	79.2	91.5	64.3	86.1	90.8	80.2	74.2	89.4	55.0	62.4	83.9	30.5	75.7	90.3	56.7	72.4
Co-SPY	77.0	91.1	59.9	92.5	93.8	91.0	88.6	90.3	86.5	90.6	93.2	87.5	79.1	89.9	65.5	87.3	88.7	85.4	85.9
D^3	62.4	82.4	31.5	71.6	86.3	51.4	82.9	90.9	73.1	80.0	88.7	68.7	82.4	89.8	73.1	73.7	84.1	57.4	75.5
Effort	64.7	93.2	31.7	**94.8**	96.3	93.2	89.7	97.1	81.9	89.5	96.0	82.3	92.9	96.2	89.2	88.0	95.0	80.2	86.6
Qwen2.5-VL-7B	50.5	64.7	2.2	56.7	90.8	15.2	50.7	70.0	2.4	53.6	90.3	8.1	54.5	90.4	10.1	51.6	63.6	4.7	52.9
InternVL3-8B	62.9	87.7	30.0	54.2	77.1	12.0	62.9	86.9	30.5	63.6	92.1	29.7	54.8	80.1	12.7	67.7	96.3	35.8	61.0
MiMo-VL-7B	48.7	47.7	26.2	82.6	76.4	**94.3**	76.4	75.1	78.9	79.7	76.3	85.7	78.4	75.8	83.5	82.8	79.1	88.3	74.8
GLM-4.1V-T	54.3	89.8	9.7	68.4	94.7	39.0	63.3	95.7	27.8	65.7	96.5	32.6	55.1	87.4	11.9	81.0	97.9	62.8	64.6
GPT-4o	66.4	82.6	41.5	58.9	**100.0**	18.0	52.5	91.6	5.5	64.4	**100.0**	29.0	60.9	94.0	23.5	55.5	**100.0**	10.1	59.8
Gemini-2.5-Pro	73.7	89.7	53.5	83.3	89.7	75.1	87.4	86.0	89.3	85.5	86.2	84.5	84.7	88.4	80.0	85.6	85.3	85.8	83.4
VERITAS	90.3	**99.5**	81.0	75.7	99.5	51.5	**97.0**	**98.6**	**95.3**	**91.8**	98.9	84.5	**95.1**	**99.5**	90.6	91.7	98.6	84.6	**90.3**
MEGRA	**91.4**	95.7	**86.7**	84.4	93.5	74.1	92.3	95.0	89.3	89.4	95.3	82.9	89.9	94.6	84.7	**93.2**	93.2	**93.2**	90.1

Table 16: CD Acc./P./R. on HydraFake. Except the MEGRA row, all entries are transcribed from VERITAS (Tan et al., 2026). InfiniteYou-CD follows the main-table CD protocol. The last row is MEGRA on the full CD test set (same global validation threshold as Table 1). Avg. is the unweighted mean of the six per-method Acc. scores. **Bold**/underline: best/second-best per column, including ties.

Method	DeepFaceLab			InfiniteYou-CD			Dreamina			Hailuo AI			GPT-4o			FFIW			Avg.
	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	Acc.	P.	R.	
F3Net	57.7	54.7	89.1	78.5	72.6	91.6	55.6	55.0	60.9	68.6	63.3	88.6	66.2	63.6	75.5	66.4	64.8	71.4	65.5
UnivFD	67.4	64.1	79.2	67.3	75.9	50.5	80.5	77.2	86.3	75.2	75.4	74.8	73.3	73.6	72.7	67.5	65.4	74.5	71.9
IID	65.2	65.1	65.5	69.9	63.0	**95.9**	63.8	58.0	99.6	63.3	57.6	**100.0**	63.8	58.2	**97.8**	64.2	66.6	56.6	65.0
FreqNet	50.6	50.3	98.7	67.0	62.3	85.8	62.1	58.4	83.4	59.3	56.9	76.8	58.3	56.3	73.3	51.2	50.6	**95.0**	58.1
ProDet	58.1	54.6	95.3	82.9	81.3	85.4	71.3	71.3	71.4	75.6	71.2	86.0	66.3	69.9	57.4	74.1	73.0	76.5	71.4
NPR	52.6	52.4	56.6	73.0	82.1	58.7	76.6	80.9	69.5	62.3	70.7	42.0	50.2	50.4	20.9	46.0	46.3	50.9	60.1
AIDE	59.7	57.5	73.9	67.9	75.2	53.2	49.7	49.4	25.2	58.0	60.6	45.8	51.9	53.0	33.3	59.2	57.3	71.2	57.7
Co-SPY	67.6	63.8	81.2	80.0	85.1	72.7	82.5	80.7	85.3	74.0	75.7	70.6	79.5	78.7	80.9	64.3	64.2	64.7	74.7
D^3	69.7	70.0	69.1	74.6	78.8	67.2	78.1	72.2	91.4	70.9	67.7	79.8	80.8	75.1	92.0	64.3	66.0	58.7	73.1
Effort	64.8	59.1	96.3	82.2	75.4	95.5	61.5	57.6	86.5	66.4	60.1	97.4	53.8	52.9	69.3	74.0	68.3	89.7	67.1
Qwen2.5-VL-7B	50.7	71.7	2.4	53.6	92.8	7.9	80.2	**99.4**	60.7	67.5	**100.0**	35.1	52.5	**100.0**	5.1	50.5	56.1	4.5	59.2
InternVL3-8B	54.4	55.5	44.1	67.1	83.6	42.7	77.1	85.4	65.3	66.5	78.9	45.0	47.4	37.5	7.6	51.8	52.0	45.1	60.7
MiMo-VL-7B	57.7	56.4	68.6	75.6	73.7	79.6	79.4	73.5	91.9	70.7	68.3	76.5	67.7	67.1	69.5	54.9	54.8	56.2	67.7
GLM-4.1V-T	58.7	**82.6**	22.0	72.7	95.1	47.8	83.7	96.0	70.3	69.2	92.1	42.0	52.0	68.4	8.2	53.9	60.8	21.7	65.0
GPT-4o	49.4	41.7	2.5	62.0	**98.0**	24.5	90.7	98.8	82.4	73.7	**100.0**	47.5	58.0	94.4	17.0	52.8	59.3	17.8	64.4
Gemini-2.5-Pro	67.2	72.4	55.6	75.6	88.7	59.0	87.5	89.5	84.9	82.4	95.2	68.2	70.9	96.6	43.2	53.0	67.6	11.5	72.8
VERITAS	58.6	54.7	**100.0**	**84.1**	94.1	72.8	92.3	90.0	95.1	90.2	86.5	95.2	89.2	86.3	93.2	**78.5**	**76.1**	83.0	82.2
MEGRA	**80.1**	74.3	91.8	80.6	89.4	69.5	**96.1**	92.8	**100.0**	**91.2**	90.7	91.8	**93.2**	93.0	93.3	73.3	73.8	72.1	**85.7**

J EXPERT POOL

Table 8 lists backbones, sources and sizes. Validation comparisons identify different leading experts across forgery families, motivating FD-MoE. Table 17 reports single-expert and unconditioned-pool average accuracy; each row fits its own validation BA threshold. These scores use the same frozen checkpoints as FD-MoE (released or DeepfakeBench weights; not fine-tuned on HydraFake) and only a class-agnostic Platt map plus a validation threshold. These values correspond to the frozen-weight protocol; the Effort and UnivFD entries transcribed from VERITAS in Table 1 follow their respective HydraFake training settings. Average accuracy follows Table 1: the same calculation as that table, not the mean of the four protocol sets. Leave-one-method-out fusion assigned zero weight to EfficientNet-B4 and F3Net (redundant CNN / already-spanned frequency cue), so the official pool keeps six of the eight scored experts.

Table 17: Frozen forensic experts and unconditioned pooling (average accuracy, %). Released or DeepfakeBench weights are used without retraining on HydraFake; each row fits its own validation BA threshold. Standalone detectors in Table 1 follow their respective training settings. MEGRA is the full detection module from Table 3.

Expert	Avg.	Expert	Avg.
Xception	56.74	UCF	62.17
RECCE	61.94	UnivFD	59.90
Effort	60.64	SBI	52.14
Uniform avg. of 6			61.58
Global stack of 6			62.93
MEGRA			**91.09**

Table 18: Nested expert-count ablation (average accuracy, %). The 4-expert pool drops SBI and Xception (keeps UCF, RECCE, UnivFD, Effort); the 2-expert pool keeps RECCE and Effort (drops Xception, SBI, UCF, UnivFD). Uniform averages fit separate validation thresholds; FD-MoE variants retain the original threshold and zero selected fusion weights while the factorized head still receives all six logits.

Experts	Uniform avg.	MEGRA
6 (all)	61.58	**91.09**
4 (drop SBI, Xception)	63.53	90.54
2 (keep RECCE, Effort)	64.02	75.92

J.1 Mixture coefficients and fusion weights

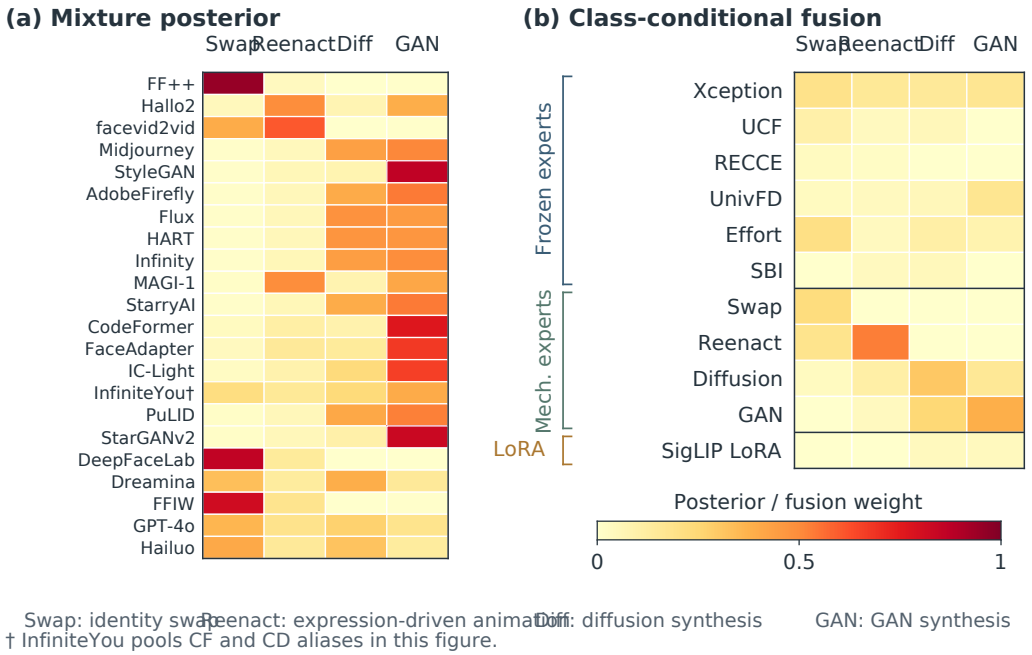


Figure 5: Mixture coefficients and fusion weights. Left: mean mixture coefficients $\tilde{p}$ by source method. Right: class-conditional fusion weights $w_{c,h}$ for the four detection-side mixture classes in $\mathcal{C}$. Reenact denotes the expression-driven animation class. The LoRA row is the SigLIP-LoRA fusion channel. †InfiniteYou pools the CF and CD protocol aliases in this figure.

K Prompts

This section documents the prompt templates used in the explanation experiments. Placeholders appear as {...}; all multimodal inputs also include the face image. Categories A–B reproduce the public VERITAS evaluation templates without modification. English wording in Listings 1–2, including "an AI assistants", "Noticement", and "evaluate score", is left unchanged from those templates. Categories C–F contain the GraphRAG student, teacher, baseline, and repair prompts. The judge models differ from VERITAS: we use GPT-5.5, Kimi-K3, and Qwen3.7-Plus rather than GPT-4o and Gemini-2.5-Pro.

Table 19: Prompt inventory. A–B are transcribed from VERITAS; C–F are ours.

Cat.	Role	Usage
A	Score judge	1–5 absolute explanation quality (GPT-5.5 / Kimi-K3 / Qwen3.7-Plus)
B	Pairwise judge	A vs. B preference → ELO rating (Kimi-K3 / Qwen3.7-Plus)
C	Student	SFT training and GraphRAG inference (Qwen2.5-VL-7B + QLoRA)
D	Teacher	Kimi reference targets for SFT supervision
E	Baseline	Qwen2.5-VL-7B inference without GraphRAG or QLoRA
F	Repair	Appended to the student user prompt on contract-validation failure

K.1 Score Evaluation (Category A)

Given ground-truth authenticity (`Real` or `Fake`), the user instruction, one assistant explanation, and the image, the judge returns JSON with keys `analysis` and `judgment` $\in$ `{[[1]],...,[[5]]}`. All three pointwise judges share this template.

Listing 1: Score-evaluation judge prompt (VERITAS template, unchanged).

```
You are a helpful assistant proficient in analyzing vision reasoning
problems.

## Instruction: Please examine the provided image attentively and serve
as an unbiased judge in assessing the quality of the response from an
AI assistants regarding the instruction. You will receive a single
response from the assistant to user's instruction.

## Noticement: Your assessment should identify whether the assistant
effectively adheres to the user's instructions and addresses the user's
inquiry.
In your evaluation, weigh factors such as preciseness,
comprehensiveness, clarity, creativity, and the granularity of the
responses.
Do not allow the length of the responses to influence your evaluation.
Do not favor certain names or positions of the assistants. Be as
objective as possible.

## Criteria:

Use scores to show the quality of the response. Here is the detailed
scoring rubric for evaluating the quality of responses from AI
assistants:
Incorrect (1): The answer is incorrect (e.g., predicting real for the
fake image), the score should be 1.
Poor (2): The response fails to address the query effectively. Although
the answer is correct, the reasoning process is vague and repetitive.
Creativity and granularity are absent or poorly executed.
Average (3): The response shows superficial analysis of the given
image. It lacks depth in creativity and granularity. The reasoning
process is not detailed.
Good (4): The response is well-aligned with the given image,
demonstrating a high degree of preciseness and comprehensiveness. It
shows creativity and a nuanced understanding of the topic, with a
detailed granularity that enhances the response quality.
Excellent (5): The response shows great preciseness, comprehensiveness,
creativity, and granularity. It provides an insightful, detailed, and
thorough answer, indicating a deep and nuanced understanding of the
user's inquiry.
```

```
## Desired Output Format:

Present your verdict in a JSON format, with the key 'analysis' for a
short reason of your judgment and the key 'judgment' to indicate your
decision: use "[[1]]", "[[2]]", "[[3]]", "[[4]]", "[[5]]" to indicate
your evaluate score.

## Question:

This is a {ground_model_output}.
[The Start of User Instruction]
Please determine the authenticity of this image.
[The End of User Instruction]
[The Start of Assistant's Answer]
{model_reasoning_output}
[The End of Assistant's Answer]
```

K.2 Pairwise Comparison (Category B)

The judge outputs `[[A]]`, `[[B]]`, or `[[C]]` (tie).

Listing 2: Pairwise comparison judge prompt (VERITAS template, unchanged).

```
You are a helpful assistant proficient in analyzing vision reasoning
problems.
## Instruction:
Please examine the provided image attentively and serve as an unbiased
judge in assessing the quality of responses from two AI assistants
regarding the user's question shown beneath the image.
## Noticement:
Your assessment should identify the assistant that more effectively
adheres to the user's instruction and provides more detailed, more
precise and high-quality reasoning.
In your evaluation, weigh factors such as preciseness,
comprehensiveness, clarity, creativity, and the granularity of the
responses.
Avoid any position biases and ensure that the order in which the
responses were presented does not influence your decision.
Do not allow the length of the responses to influence your evaluation.
Do not favor certain names of the assistants. Be as objective as
possible.
## Desired Output Format:
Present your verdict in a JSON format, with the key 'analysis' for a
short reason of your judgment and the key 'judgment' to indicate your
decision: use "[[A]]" if assistant A prevails, "[[B]]" if assistant B
does, and "[[C]]" for a tie.
## Question:
This is a {ground_model_output}.
[The Start of User Instruction]
Please determine the authenticity of this image.
[The End of User Instruction]
[The Start of Assistant A's Answer]
{assistant_a_output}
[The End of Assistant A's Answer]
[The Start of Assistant B's Answer]
{assistant_b_output}
[The End of Assistant B's Answer]
```

K.3 GraphRAG Student (Category C)

SFT and inference share the student input and output requirements. The authenticity decision is supplied as `fixed_decision` in `AVAILABLE INPUT` and copied into

the output field `upstream_final_decision`, implementing Eq. (2). The structured `forgery_type` field uses the explanation vocabulary (`face_swap`, `reenactment`, `entire_face_generation`). Runtime fields are serialized as JSON.

Listing 3: Student system prompt.

```
[System]
You explain a fixed face-authenticity decision; you do not classify the
image again.
Inspect the supplied image and use the candidate regions and cue
hypotheses only as places to verify, not as facts. Select only details
that are visibly supportable. Conditional forensic knowledge may help
interpret an observation, but it must never be repeated as
implementation language or treated as proof.
Write one natural, reader-facing forensic explanation. Describe how a
visual change appears, compare it with neighboring or global content,
explain why multiple observations are related, and calibrate ordinary
alternatives such as compression, blur, pose, lighting, or capture
quality. For a real decision, give positive continuity evidence rather
than merely saying that no artifact exists.
For a fake decision, predict exactly one broad forgery type: face_swap,
reenactment, entire_face_generation, or unknown. This is a type
judgment, not a claim about a specific implementation or generator. For
a real decision, the structured forgery_type is unknown and the prose
has no forgery-type closing sentence.
Return exactly one raw JSON object. Do not use Markdown fences or text
outside the JSON. The prose must not mention GraphRAG, retrieval,
paths, chunks, nodes, rules, scores, fields, datasets, files,
source-method metadata, or generator identities.
```

Listing 4: Student user prompt template. Braced fields are filled at runtime. The phrase "fixed decision" denotes `fixed_decision`.

```
[User]
Explain the fixed decision for the supplied image.

PART-EVIDENCE TOKENS
{part_evidence_tokens}

AVAILABLE INPUT
{available_input_json}

ALLOWED REGIONS
{allowed_regions}

ALLOWED CUES
{allowed_cues}

OUTPUT SCHEMA
{output_schema_json}

WRITING REQUIREMENTS
- Copy the fixed decision exactly. Do not reclassify the image.
- Select one to three visibly supportable regions and cues. Candidate
ordering is only an inspection hint, not evidence strength; do not list
every candidate.
- Usually write 150-220 English words, but prioritize complete,
non-repetitive reasoning over a fixed length.
- Describe concrete appearance and an adjacent or global comparison; do
not merely translate cue names.
- Explain why observations support one another instead of calling them
coordinated or complementary.
- Use ordinary quality factors as calibrated alternatives, not as
forgery evidence.
```

```
- For fake, predict exactly one broad forgery_type from face_swap,
reenactment, entire_face_generation, or unknown. Use unknown only when
the image and evidence cannot support a broad type.
- End every fake explanation with exactly the matching sentence from
this map: {forgery_type_tails_json}
- For real, set forgery_type=unknown, provide positive visual
continuity, and do not add any forgery-type closing sentence.
- Do not list two candidate types and do not identify a specific
implementation, editing pipeline, or generator.
- Return raw JSON only.
```

Table 20: Main fields in `AVAILABLE INPUT`. The student copies `fixed_decision` into output `upstream_final_decision`. The output schema also requires `evidence_regions`, `visual_cues`, `forgery_type`, `uncertainty`, and `human_explanation`.

Field	Meaning
`fixed_decision`	Authenticity decision (`real` / `fake`); copied verbatim into `upstream_final_decision`
`confidence_state`	Confidence bucket (`low|medium|high`)
`candidate_regions`	Canonical region hypotheses
`cue_hypotheses`	Canonical cue hypotheses
`conditional_forensic_knowledge`	Cue-linked interpretive hints
`interpretation_boundary`	Candidates are inspection hints, not observed facts

K.4 TEACHER (CATEGORY D)

Kimi writes reference explanations from grounded visual facts. A style-reference block with positive forensic-reasoning exemplars and negative phrasing constraints is prepended, inspired by contrastive in-context learning (Gao & Das, 2024); the full style list is omitted here.

Listing 5: Teacher system and user prompts (style block abbreviated).

```
[System]
You are the teacher writer for a face-forensics explainer. The final
Real/Fake decision and supervised broad forgery type are fixed training
targets. The supplied grounded facts are reliable visual anchors, not a
closed whitelist. Inspect the image itself: you may add compatible,
visibly supportable detail in neighboring, symmetric, transitional, or
global regions, but must not invent an independent fact that cannot be
seen. Conditional knowledge explains how cues may be interpreted; it is
not an observed fact.
Write a natural forensic analysis. [...]
Return one raw JSON object only.

[User]
Create the teacher target for this sample.
{STYLE_REFERENCE}
SOURCE MATERIAL
{source_material_json}
SAMPLE-SPECIFIC WRITING EMPHASIS
{sample_specific_writing_emphasis}
OUTPUT SCHEMA
{output_schema_json}
```

K.5 QWEN2.5-VL-7B BASELINE (CATEGORY E)

Open-model control without GraphRAG retrieval or QLoRA.

Listing 6: Qwen2.5-VL-7B baseline inference prompts.

```
[System]
You are an image authenticity expert. Your task is to determine the
authenticity of the given facial image.

Firstly, give an overall judgment to the authenticity of the image,
enclosed in <fast> </fast> tags.
Then, make a careful and structured thinking before reaching an answer.
Based on your thinking, draw a comprehensive conclusion. Enclose the
corresponding part in different tags, e.g., <planning> or <reasoning>
or <reflection> or <conclusion>.
Finally, give the final answer with "real" or "fake", enclosed in
<answer> </answer> tags.

[User]
Please determine the authenticity of this image.
```

K.6 Contract Repair (Category F)

If the first student JSON fails schema or policy checks, this paragraph is appended and the model regenerates once.

Listing 7: Repair suffix appended to the student user prompt.

```
The previous answer failed these output checks: {diagnostic}.
Regenerate one complete raw JSON object only. Preserve the fixed
Real/Fake decision, use only regions and cues supplied in this prompt,
and keep uncertainty proportional to the available evidence. Do not add
Markdown fences or any text outside the JSON. Rewrite the explanation
as natural image analysis rather than describing fields, retrieval,
paths, rules, scores, or the generation process.
```

K.7 Controlled vocabulary

Student and teacher JSON use the output-contract vocabulary in Table 21: nine regions and twelve cues, all as snake_case tokens (e.g. `face_boundary`). This vocabulary is distinct from the eight detection-module seed regions and the graph ontology of 21 regions and 27 cues (Appendix D); it is not the full graph inventory. Fake explanations end with exactly one sentence from Table 22.

Table 21: Output-contract region and cue vocabularies injected into student and teacher prompts.

Allowed regions	Allowed visual cues
`eyes`, `mouth`, `nose`, `skin`, `jawline`, `face_boundary`, `hairline`, `eyebrows`, `background`	`texture_inconsistency`, `color_mismatch`, `boundary_blending`, `lighting_inconsistency`, `geometry_distortion`, `identity_inconsistency`, `expression_inconsistency`, `unnatural_eyes`, `unnatural_mouth`, `compression_ambiguity`, `low_quality_ambiguity`, `no_consistent_forgery_pattern`

Table 22: Broad forgery-type closing sentences appended to fake explanation prose.

`forgery_type`	Required closing sentence (fake only)
`face_swap`	The most likely forgery type is face swapping.
`reenactment`	The most likely forgery type is face reenactment.
`entire_face_generation`	The most likely forgery type is entire-face generation.
`unknown`	The forgery type remains unknown.

L JUDGE PROTOCOL

Score averages use successful judgments; gains use unrounded means. Three judges evaluate 1,000 randomly sampled images (300 real / 700 fake). Each pairwise comparison has 1,000 images: Kimi-K3 succeeds on 1,000 for the VERITAS pairs and Qwen3.7-Plus on 998; other MEGRA-versus-X pairs differ by at most one call. Pairwise judges return win, loss, or tie. We convert decided pairwise outcomes to ELO ratings; ties are excluded. For a pair, the empirical win rate and rating gap are

$$P = \frac{W}{W + L}, \qquad \Delta = 400 \log_{10} \frac{P}{1 - P},$$

where W and L are win counts and Δ is the rating gap on a scale of 400. If either $W = 0$ or $L = 0$, both counts are Laplace-smoothed as $W+0.5$ and $L+0.5$ before computing Δ; otherwise the raw decided counts are used. Table 2 uses one global fit per judge over all MEGRA-versus-X pairings (baselines are not paired with each other), mean-anchored at 1,000. Because the comparison graph is star-shaped with MEGRA as the hub, this global fit recovers the pairwise ELO gaps above. Qwen3.7-Plus ELO for the Qwen2.5-VL-7B base is omitted there due to suspected same-family preference; other Qwen3.7-Plus ELO entries keep that global fit. Input-ablation ELO uses the same formula in a separate three-method pool and is reported separately from Table 2 (Kimi-K3: 1,000 successes; Qwen3.7-Plus: 999). No human evaluation is reported.

M REPRODUCIBILITY ARTIFACTS

A single test pass with the frozen detection module yields the reported average accuracy of 91.09%. The explanation model is the QLoRA-adapted Qwen2.5-VL-7B checkpoint used in §4.1 and Tables 2–4. Anonymous supplementary code includes training and evaluation scripts, the average-accuracy subset mapping, and category maps; it does not alter the frozen detection-module predictions.

N ADDITIONAL QUALITATIVE EXAMPLES AND LIMITATIONS

N.1 BASELINE RESPONSE ANALYSIS

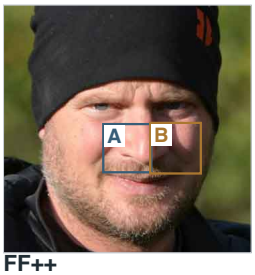
The other baselines show distinct limitations in the archived responses and judge rationales. M2F2-Det (1.55/1.77 under Kimi-K3/Qwen3.7-Plus) often substitutes scene descriptions for authenticity reasoning; low-scored outputs include missing or incorrect verdicts, bare assertions, and repetition. SIDA-7B (1.62/1.59) produces only three distinct class-template responses in this run, offering little image-specific explanation even when the class is correct; Kimi-K3 assigns score 1 to 509 of 1,000 responses. FakeVLM scores higher (2.95/2.33), but judges flag repeated generic rationales and unsupported details, such as references to body parts outside a face crop.

N.2 QUALITATIVE EXAMPLES

Figures 6–15 show 20 qualitative examples (10 real and 10 fake), two cases per page. They use the same explanation-module checkpoint as Figure 3. Green highlights mark manually inspected visual correspondences and, on fake examples, the explanation module's correct type sentences; lettered crops assist the inspection of visual correspondences. The fake examples include face swapping, reenactment, and entire-face generation (diffusion-based and GAN-based).

Figure 16 shows two cases in which the binary fake decision is correct but parts of the explanation are unreliable. Both cases retain useful local descriptions highlighted in green, while their red type sentences disagree with the source-method categories (face swapping and reenactment, respectively). In the second case, the red text also dismisses makeup or denoising on the unsupported premise that they cannot leave skin smoother than the background. These cases show that accurate detection and useful local observations do not guarantee reliable reasoning or forgery-type attribution throughout an explanation.

(a) Case 01 **MEGRA** FF++ Real



FF++
Reference: real



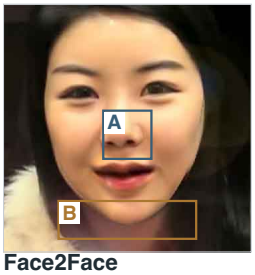
A Skin



B Skin

The image is assessed as real. The strongest support comes from the way illumination behaves across the frame: shadows on the face fall in directions and strengths that match the lighting visible in the background, so the subject reads as photographed under a single scene rather than inserted into it. That agreement extends to the skin itself, where tonal gradation stays even and there are no abrupt color steps or texture mismatches between the face and its surroundings. Because the background shares the same light treatment as the subject, the two regions read as parts of one continuous exposure instead of separately lit elements. The medium overall quality deserves mention: ordinary compression, mild blur, and low resolution could soften fine detail, but such effects act uniformly across the frame and do not produce the localized discontinuities that manipulation tends to leave behind. No region shows a boundary where content looks pasted over its neighbors, and no cue points to any specific editing method. Consistency of this kind cannot prove authenticity absolutely, yet nothing in the visible evidence contradicts an authentic capture.

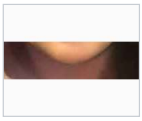
(b) Case 02 **MEGRA** Face2Face Fake



Face2Face
Reference: reenactment



A Nose

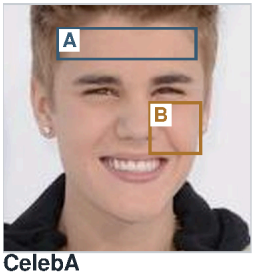


B Jawline

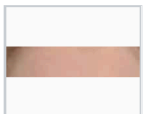
The image is assessed as fake. The decisive pattern is not a single odd patch but a repeated local mismatch across the face: skin tone shifts slightly from one area to the next, and the jawline reads as a softened contour rather than a continuous edge that carries the same sharpness as the surrounding facial structure. In a genuine capture, skin tone varies gradually with illumination and focus; here the tonal change is confined to small patches rather than following any plausible lighting or depth gradient. The jawline is the same kind of boundary where a composited region would most often betray itself, yet it behaves like a surface feature whose edges have been smoothed instead of a contour whose outline was cut out and reinserted. Medium overall quality leaves room for ordinary compression or denoising to soften isolated details, and makeup or defocus could account for part of the appearance in isolation. What those factors cannot explain is why the tonal shift and the jawline softening occur together while neighboring skin retains different detail. The evidence supports the fake decision with high confidence, though no specific manipulation method can be identified from these observations alone. The most likely forgery type is face reenactment.

Figure 6: Qualitative explanations for cases 01–02. Green highlights image-specific descriptions and the explanation module's correct type sentences.

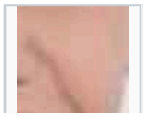
(a) Case 03 **MEGRA** CelebA Real



CelebA
Reference: real



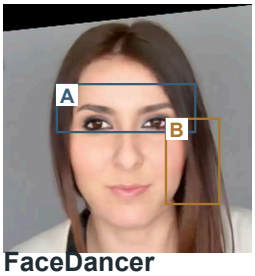
A Hairline



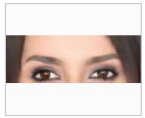
B Face boundary

The image is assessed as real. The strongest support comes from the way neighboring structures hold together rather than from any single detail. Skin tone stays even across the face, and the hairline follows a natural contour that matches the surrounding skin without a softened rim or abrupt color step where the head meets the background. That continuity matters because a composited or regenerated face typically betrays itself at exactly this junction: a pasted edge would be expected to show a blurred or mismatched boundary against the adjacent hair or background. Here the transition reads as one continuous surface, with no localized softening or tonal break. The frame is of medium quality, so ordinary compression and mild blur could account for some loss of fine texture in isolated areas; such effects act broadly across the whole image and do not produce the localized discontinuities that manipulation usually leaves behind. Because the visible evidence shows compatible structure across neighboring regions instead of any localized anomaly, the assessment is conservative but not absolute — visual consistency cannot prove authenticity in an absolute sense, yet nothing in the inspected content contradicts it.

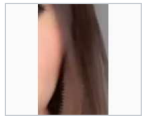
(b) Case 04 **MEGRA** FaceDancer Fake



FaceDancer
Reference: face swapping



A Eyes

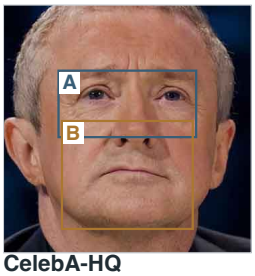


B Face boundary

The image is assessed as fake. The strongest signal sits in the central face: the eyes are not aligned with one another, so the features read as though they belong to different parts of a head rather than to a single head. That misalignment is not a minor asymmetry; it is a **structural mismatch** across the same facial structure that should be proportionate and symmetric. [...] The cheeks and forehead show a slightly different grain from the surrounding skin, and the **jawline and hairline sit at a different sharpness level than the adjacent skin**. [...] Pose or lens perspective can shift apparent eye spacing, and compression or denoising can soften texture. What those factors cannot easily explain is a **local texture break that coincides with a global geometric failure**. The most likely forgery type is **face swapping**.

Figure 7: Qualitative explanations for cases 03–04. Green highlights image-specific descriptions and the explanation module's correct type sentences.

(a) Case 05 **MEGRA** CelebA-HQ Real



CelebA-HQ
Reference: real



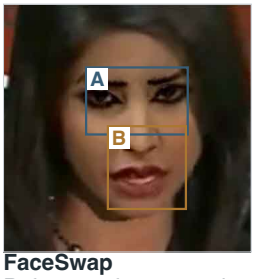
A Eyes



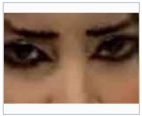
B Mouth

The image is assessed as real. The strongest support comes from the way the face meets its surroundings: the hairline and the outer contour of the head carry a continuous tonal gradient that matches the skin immediately beside it, with no abrupt change in sharpness or color where the face meets the background. That boundary behavior is what compositing or re-rendering typically disrupts first, because those operations usually leave a seam or a mismatched edge; here the transition reads as one continuous photographic surface rather than a pasted one. The interior of the face reinforces this reading. Skin tone stays even across the cheeks and forehead, and the expression sits naturally on the features, so the face does not appear to have been reshaped or re-expressed after capture. Because the frame is only medium quality, ordinary compression and mild blur could soften isolated details, but such effects act uniformly across the whole image and would not selectively erase the boundary continuity observed here. No localized anomaly stands out against this coherent account, and the evidence supports a confident real determination while acknowledging that visual consistency alone cannot prove authenticity absolutely.

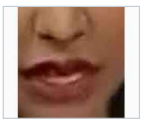
(b) Case 06 **MEGRA** FaceSwap Fake



FaceSwap
Reference: face swapping



A Eyes



B Mouth

The image is assessed as fake. The strongest signal sits in the eyes: their shape and spacing do not sit naturally with the surrounding facial structure, so the eyes read as if they were placed onto a head rather than grown from it. That local mismatch matters more than ordinary asymmetry because the rest of the face is otherwise coherent; skin tone stays even across the frame, and the expression reads as one connected state rather than a pasted or reshaped feature. The nose reinforces this reading. Its proportions are slightly off and its surface texture does not match the neighboring skin, which is exactly where recomposition tends to leave a seam. The mouth area shows a similar problem: its shape and expression do not agree with the rest of the face. Because these deviations appear at three separate facial landmarks — the eyes, the nose, and the mouth — they describe a single failure mode rather than isolated quirks. Medium overall quality leaves room for compression, defocus, or makeup to soften detail, but those effects would be expected to act broadly across the frame and would not selectively alter eye geometry while leaving the rest of the face intact. The most likely forgery type is face swapping.

Figure 8: Qualitative explanations for cases 05–06. Green highlights image-specific descriptions and the explanation module's correct type sentences.

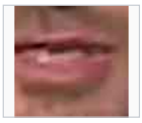
(a) Case 07 **MEGRA** FF++ Real



FF++
Reference: real



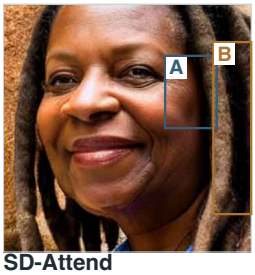
A Eyes



B Mouth

The image is assessed as real. The strongest support comes from the way neighboring structures hold together as one coherent structure: the eyes sit symmetrically and naturally with respect to the nose, and the spacing between them agrees with the width of the face rather than appearing inserted or reshaped. That geometric agreement matters because a pasted or regenerated face usually betrays itself first at exactly this kind of alignment — the eyes that should be aligned with the nose instead look placed off-center or mismatched in scale. Here the spacing is consistent across the whole face, so the eye-nose relationship reads as a single anatomical unit rather than a composite. The skin reinforces the same account. Tone stays even across the face, and the expression sits naturally on the features it moves, with no localized warping or mismatched tension around the eyes or mouth. Because the medium overall quality means ordinary compression or mild blur could soften isolated details, those effects are not decisive; they would act broadly across the frame and cannot manufacture the cross-region agreement observed here. Consistency of this kind supports authenticity strongly but does not prove it absolutely.

(b) Case 08 **MEGRA** SD-Attend Fake



SD-Attend
entire-face generation



A Face boundary

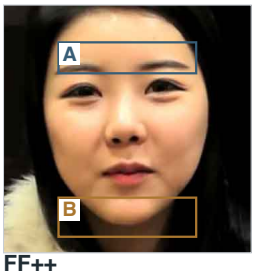


B Background

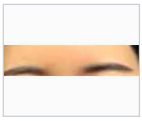
The image is assessed as fake. The strongest signal sits at the face's outer edge: the transition from skin to background reads as a softened seam rather than a gradual tonal falloff, and the hairline and jawline follow the same flattened logic. That local discontinuity matters because the rest of the frame is of medium quality but still coherent; blur and low resolution are present across the whole image, so they would be expected to soften everything uniformly. Here the loss of fine detail concentrates on the facial boundary while the interior keeps its character, which is what an inserted or synthesized face tends to leave behind when it is not fully integrated into the scene. The background reinforces this reading. It is unusually sharp relative to the subject, and the lighting on the face does not match the illumination falling on the backdrop. These two observations belong together because they describe one connected event: a face placed into a scene that was not captured under the same conditions. Makeup, defocus, compression, or uneven exposure could each explain part of any single detail in isolation, but they do not easily account for a localized texture break at the face boundary combined with a mismatched background. The most likely forgery type is entire-face generation.

Figure 9: Qualitative explanations for cases 07–08. Green highlights image-specific descriptions and the explanation module's correct type sentences.

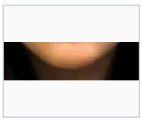
(a) Case 09 **MEGRA** FF++ Real



FF++
Reference: real



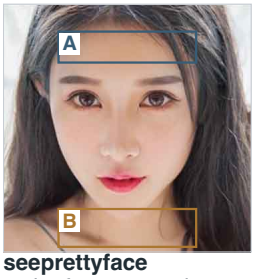
A Hairline



B Jawline

The image is assessed as real. The strongest support comes from the way neighboring structures hold together: the hairline meets the face boundary with a smooth transition rather than a softened rim or a blurred seam, and the same continuity continues along the jawline where it meets the neck. In a composited or regenerated face, that kind of edge behavior is precisely what tends to betray manipulation—where a new region was inserted, the blending would usually leave a faint halo or a mismatch in sharpness at the contour. Here there is no such discontinuity; the facial outline reads as one continuous surface extending into the surrounding hair and neck. The medium overall quality deserves mention: ordinary compression, mild blur, and lighting variation could account for some softening in isolated areas, but those effects act broadly across the frame and do not produce the localized discontinuities that editing typically leaves behind. The evidence therefore supports authenticity strongly without proving it absolutely.

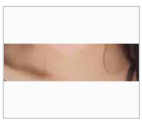
(b) Case 10 **MEGRA** seeprettyface Fake



seeprettyface
entire-face generation



A Hairline

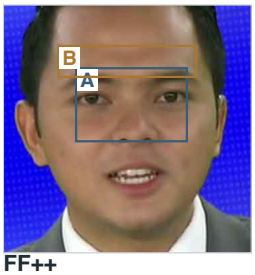


B Jawline

The image is assessed as fake. The decisive evidence sits at the face's outer boundary: the jawline and hairline show a slight mismatch in sharpness against the skin immediately beside them, so the contour reads as slightly softened rather than as a continuous edge. That local softening is not matched by any comparable change elsewhere on the face; the skin itself remains uniformly smooth with no visible pore-level detail, and the expression appears natural. Because the jawline and hairline are exactly where composited or regenerated content tends to sit relative to surrounding skin, their localized softness is harder to dismiss as ordinary capture or compression. Medium overall quality leaves room for ordinary causes such as denoising or makeup to contribute to the smoothness of the skin, but those effects would be expected to act broadly across the frame rather than concentrate along the facial perimeter. The most likely forgery type is entire-face generation.

Figure 10: Qualitative explanations for cases 09–10. Green highlights image-specific descriptions and the explanation module's correct type sentences.

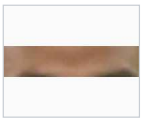
(a) Case 11 **MEGRA** FF++ Real



FF++
Reference: real



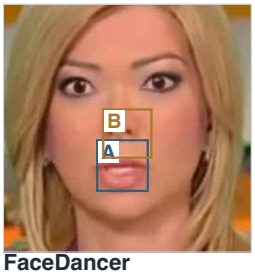
A Eyes



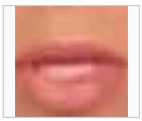
B Hairline

The image is assessed as real. The strongest support comes from the way neighboring structures hold together as one photographic account: the eyes sit in proportionate alignment with the surrounding face, and the hairline follows a natural contour that continues into the forehead without a softened rim or abrupt change in texture. That continuity matters because compositing or re-rendering typically leaves a boundary where the inserted content meets its surroundings; here the transition reads as ordinary capture rather than a seam. The medium overall quality deserves mention, since it affects the frame uniformly and does not concentrate on any single region. Compression or mild blur could soften fine detail, but such effects would act across the whole image and would not selectively erase the eye-to-face geometry or the hairline's smoothness. No localized artifact of this kind is visible. The evidence therefore supports authenticity strongly while stopping short of absolute proof, since visual consistency alone cannot exclude every conceivable manipulation.

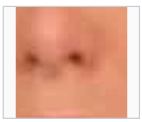
(b) Case 12 **MEGRA** FaceDancer Fake



FaceDancer
Reference: face swapping



A Mouth

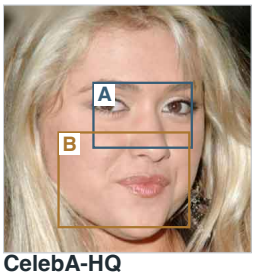


B Nose

The image is assessed as fake. The decisive pattern is not a single odd feature but a repeated local mismatch across the face: the skin reads as unusually smooth in places where neighboring skin retains more of its natural grain, and the mouth area shows a slightly different texture from the surrounding cheek skin. Because the frame is only medium quality with visible blur and low resolution, ordinary compression or denoising could soften detail broadly; what matters here is that the change stays confined to specific patches rather than spreading evenly across the whole face. A global degradation would be expected to reduce sharpness uniformly, while here the smoothing concentrates on isolated regions. Makeup or heavy retouching could also explain part of this appearance, so it cannot be dismissed outright. The most informative observation is that the texture breaks exactly at the boundaries between facial structures — the jawline and the mouth — which is where recomposition tends to leave residual seams. The evidence supports the fake decision with high confidence, though no specific manipulation method can be identified from these observations alone. The most likely forgery type is face swapping.

Figure 11: Qualitative explanations for cases 11–12. Green highlights image-specific descriptions and the explanation module's correct type sentences.

(a) Case 13 **MEGRA** CelebA-HQ Real



CelebA-HQ
Reference: real



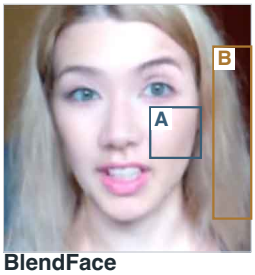
A Eyes



B Mouth

The image is assessed as real. The strongest support comes from the way illumination behaves across the frame: shadows on the face fall in directions that agree with the lighting visible in the background, and the subject's expression reads as a coherent whole rather than a frozen or mismatched feature. That agreement matters because a composited or re-rendered face typically betrays itself through a mismatch between the subject's shading and the scene behind it; here the two regions share one consistent light source. Skin tone stays even across the face, and the facial features sit in proportionate alignment with one another, so no single region looks pasted onto its neighbors. The medium overall quality deserves mention: ordinary compression, mild blur, and flat lighting could each account for isolated softness or tonal variation, but they act uniformly across the frame rather than concentrating along any contour or feature boundary. No localized discontinuity stands out against this backdrop. The evidence therefore supports a confident real judgment without proving authenticity absolutely.

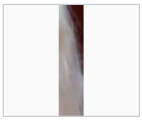
(b) Case 14 **MEGRA** BlendFace Fake



BlendFace
Reference: face swapping



A Skin

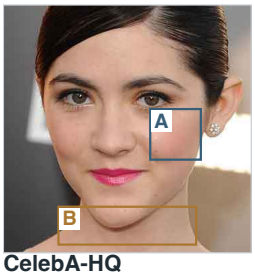


B Face boundary

The image is assessed as fake. The decisive evidence sits at the skin surface: across the face the texture reads as unusually smooth, lacking the pores and fine lines that normally survive even moderate compression or defocus. That smoothing is not uniform across the frame; it concentrates on the skin while the background retains a different level of grain and sharpness. A global capture effect would be expected to soften everything together, yet here the skin is the only area where detail is lost relative to its surroundings. The medium overall quality leaves room for ordinary explanations. Compression can flatten texture, and denoising or makeup can suppress fine detail, so a single smoothed patch could plausibly stem from those causes. What resists that account is the selective nature of the change: the loss of skin microstructure is confined to the facial surface while neighboring content keeps its own character. That local selectivity is what makes the observation informative rather than merely a statement that the whole image is degraded. The most likely forgery type is face swapping.

Figure 12: Qualitative explanations for cases 13–14. Green highlights image-specific descriptions and the explanation module's correct type sentences.

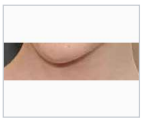
(a) Case 15 **MEGRA** CelebA-HQ Real



CelebA-HQ
Reference: real



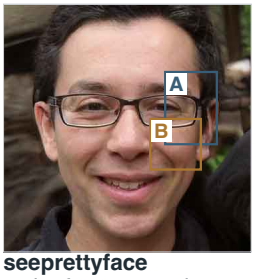
A Skin



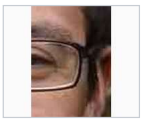
B Jawline

The image is assessed as real. The strongest support comes from the way illumination behaves across the face: shadows fall in a physically plausible direction, and the skin tone stays even from one region to the next rather than breaking at any contour. That continuity matters because compositing or re-rendering typically leaves a seam where the inserted content meets its surroundings, and such seams usually betray themselves through a mismatch in shading or color. Here there is no such break along the facial outline or between the skin and the background; the lighting reads as one coherent capture instead of a pasted element. The medium overall quality deserves mention: ordinary compression, mild blur, and low resolution could each soften isolated details, but they act uniformly across the frame and cannot manufacture the consistent shadow behavior observed. Because several neighboring regions preserve compatible structure, the evidence supports a confident real judgment without proving authenticity absolutely.

(b) Case 16 **MEGRA** seeprettyface Fake



seeprettyface
entire-face generation



A Face boundary

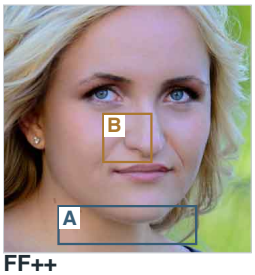


B Skin

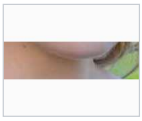
The image is assessed as fake. The strongest signal sits along the outer contour of the face: the edge where the skin meets the background reads slightly sharper than the surrounding scene, and the texture immediately inside that boundary is smoother than what appears on neighboring skin. A genuine capture would show a gradual transition from face to background rather than a localized step in sharpness and surface detail. The interior of the face tells a similar story. Skin tone stays even across the visible surface, and the expression looks natural, so the problem is not a global color mismatch or an expression that does not fit the rest of the face. What matters here is that the change is confined to the boundary while the rest of the frame keeps its own character; uniform compression or defocus would soften everything together instead of leaving the edge sharper than its surroundings. Makeup or denoising could plausibly account for part of this appearance, but they would not selectively sharpen one contour relative to adjacent content. The most likely forgery type is entire-face generation.

Figure 13: Qualitative explanations for cases 15–16. Green highlights image-specific descriptions and the explanation module's correct type sentences.

(a) Case 17 **MEGRA** FF++ Real



FF++
Reference: real



A Jawline



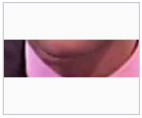
B Nose

The image is assessed as real. The strongest support comes from the way neighboring structures agree with one another: the jawline and nose share a continuous geometric relationship, so the transition between them reads as a single coherent contour rather than two separate regions that have been assembled together. That same continuity extends into the skin itself, where tone and texture remain even across the face without the patchiness or abrupt tonal steps that often appear when a region has been re-rendered or blended. The expression reinforces this reading, because the mouth and surrounding skin move together in a plausible configuration instead of showing the mismatched tension that manipulation tends to leave behind. The frame is only medium quality, which means ordinary compression and mild blur could soften isolated details, but those effects would act broadly across the whole image and cannot manufacture the consistent geometry and texture observed here. No localized anomaly stands out against this account, so the evidence supports a confident real determination while acknowledging that visual consistency alone cannot prove authenticity absolutely.

(b) Case 18 **MEGRA** Face2Face Fake



Face2Face
Reference: reenactment



A Jawline

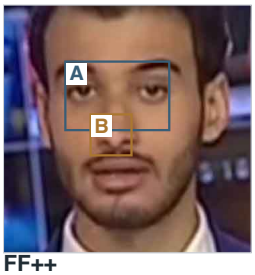


B Face boundary

The image is assessed as fake. The decisive evidence sits at the jawline and the surrounding skin: along that contour the surface reads slightly differently from the skin immediately beside it, so the boundary does not behave like a single continuous surface. A genuine face shows gradual tonal change across the cheek and neck, with fine detail softening gradually rather than stopping abruptly against the jaw; here the transition is abrupt instead of gradual. That local discontinuity matters because the rest of the frame is of medium quality but still retains some fine detail elsewhere, so the change cannot be explained by uniform compression or defocus affecting the whole image. Makeup, denoising, or ordinary lighting could each soften texture in isolation, and pose or lens perspective could shift apparent geometry, but they would not produce an edge where neighboring skin keeps different detail. The most likely forgery type is face reenactment.

Figure 14: Qualitative explanations for cases 17–18. Green highlights image-specific descriptions and the explanation module's correct type sentences.

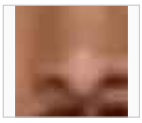
(a) Case 19 **MEGRA** **FF++** **Real**



FF++
Reference: real



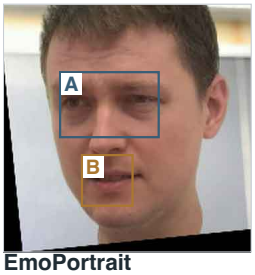
A Eyes



B Nose

The image is assessed as real. The strongest support comes from the way neighboring structures hold together as one coherent structure: the eyes sit symmetrically within a face whose proportions and spacing are natural, and the nose sits in the same proportionate relationship to the surrounding features. That mutual alignment matters because a pasted or regenerated face usually betrays itself exactly at these junctions—where the eyes meet the cheeks, where the nose meets the forehead, and where the mouth meets the cheeks—because those boundaries are precisely where inserted content tends to lose its own geometry while trying to fit the surrounding skin. Here none of those seams shows any such mismatch; the eye region reads as part of the same facial surface as the nose and the rest of the face, with no localized blur, color step, or tonal break that would suggest a recomposed boundary. The medium overall quality does soften fine detail, but that softening acts uniformly across the frame rather than concentrating along any single feature boundary. Ordinary compression or mild blur could account for some of this softness, so visual consistency alone cannot prove authenticity absolutely. Still, nothing in the inspected regions contradicts an ordinary capture explanation.

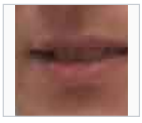
(b) Case 20 **MEGRA** **EmoPortrait** **Fake**



EmoPortrait
Reference: reenactment



A Eyes

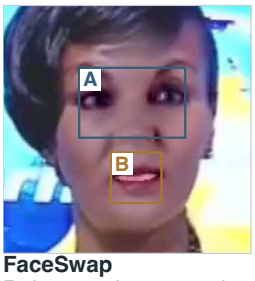


B Mouth

The image is assessed as fake. The decisive pattern is not a single patch of odd texture but a repeated local mismatch across the face: the skin around the eyes and mouth reads slightly differently from the skin immediately beside it, so the same surface area shows two distinct grain structures instead of one continuous one. That matters because the frame is only medium quality overall; compression or denoising would be expected to soften detail evenly across the whole image, and defocus would blur edges broadly rather than leave isolated patches where neighboring skin looks differently textured. Makeup or heavy retouching could plausibly account for part of this appearance in isolation, since both can smooth fine detail, but such effects tend to act uniformly on a region rather than produce the localized variation seen here. The most likely forgery type is face reenactment.

Figure 15: Qualitative explanations for cases 19–20. Green highlights image-specific descriptions and the explanation module's correct type sentences.

(a) Case 21 **MEGRA** FaceSwap Fake



FaceSwap
Reference: face swapping



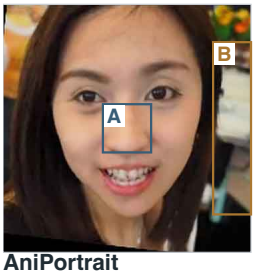
A Eyes



B Mouth

The image is assessed as fake. The strongest signal sits in the central face: the eyes and nose read as slightly misaligned, with the left eye appearing larger than the right and the nose sitting off-center relative to the rest of the face. That kind of asymmetry is difficult to explain by ordinary facial asymmetry because the deviation is not confined to one feature but spans two connected structures. Around the mouth, the skin texture changes abruptly compared to the neighboring cheek and jaw areas; the change is local rather than uniform across the frame, which is what makes it informative. Medium overall quality leaves room for compression or defocus to soften detail, and makeup or denoising can smooth texture, so no single cue identifies a manipulation method. What matters is that the texture shift stays local while the surrounding skin keeps different detail, and the geometric mismatch between the eyes and nose sits in the same region. These observations reinforce each other rather than standing alone. The most likely forgery type is face reenactment.

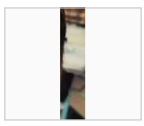
(b) Case 22 **MEGRA** AniPortrait Fake



AniPortrait
Reference: reenactment



A Nose



B Background

The image is assessed as fake. The decisive evidence sits at the skin surface rather than in any single feature: across the face the texture reads as unusually smooth, with pores and fine lines absent where a genuine capture would retain some of that micro-structure even under medium-quality conditions. That loss of detail is not uniform across the frame; it concentrates on the skin while the background retains a slightly different grain, so the change does not look like ordinary compression or defocus softening the whole image. A global degradation would be expected to flatten edges and reduce high-frequency content everywhere at once, whereas here the skin is the only region where the surface character is stripped away relative to its surroundings. Makeup or denoising could plausibly account for part of this appearance, since both can smooth skin texture, but such processing typically acts broadly and would not leave the skin visibly smoother than the background. The most likely forgery type is face swapping.

Figure 16: Explanation errors in cases 21–22 despite correct fake decisions. Green highlights useful local descriptions; red marks incorrect type attribution and an unsupported dismissal of benign processing.