COUNTERFAKE: TOWARDS GENERALIZABLE MLLM-BASED DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deepfake detection remains challenging under distribution shifts caused by unseen data domains and evolving forgery techniques. Multimodal large language model (MLLM) based detectors show promising generalization but remain vulnerable to these shifts. We introduce *CounterFake*, an MLLM-based deepfake detector designed to improve robustness to various domain and forgery shifts. To improve detection of fake images from unseen domains, we develop two complementary *counterfactual preference losses*: an image-level loss using hard-negative images and a text-level loss using counterfactual rationales. These losses are supported by *ReForensics* (Reference-guided Forensics Annotation), a comparative captioning pipeline that uses reference images to generate forensic rationales grounded in decisive forgery cues. We further highlight two simple yet important design choices for MLLM-based deepfake detection: (*i*) *weight-space interpolation* improves robustness on real images from unseen domains, while (*ii*) *intermediate representation probing* reads out authenticity information directly from intermediate language model features while retaining the ability to generate forensic rationales. Extensive experiments on DF40-based benchmarks demonstrate that CounterFake outperforms existing state-of-the-art MLLM-based deepfake detectors, *e.g.*, by 9.1 and 7.9 percentage points in average macro-F1 under cross-domain shifts and joint domain/forgery shifts, respectively.

1 INTRODUCTION

The advancement of visual generative AI (Choi et al., 2018; Karras et al., 2020; Rombach et al., 2022; Podell et al., 2023; Esser et al., 2024; Xie et al., 2024; Chen et al., 2025; Qin et al., 2025) has significantly advanced the creation and manipulation of highly realistic facial imagery. However, these advances have also amplified the risks of misuse, including the generation of deceptive facial content for identity fraud and financial scams (Cross, 2022; de Rancourt-Raymond & Smaili, 2023; Lin, 2025). As a result, reliable detection of facial deepfakes has become a critical and urgent research challenge (Amaresam, 2024; Shetye et al., 2025).

However, applying facial deepfake detection in real-world scenarios remains challenging due to substantial domain discrepancies between training and test data. Such domain shifts can significantly degrade detection performance and hinder reliable generalization beyond the training distribution (Chen et al., 2022; Wang et al., 2023; Park et al., 2025). These shifts can arise in two distinct forms: (*i*) a *real-side domain shift*, driven by variations in image content and acquisition or processing conditions, such as composition, resolution, compression, and resizing; and (*ii*) a *fake-side domain shift*, associated with differences in the source images used for manipulation and evolving forgery methods. Conventional vision-based deepfake detectors have demonstrated strong performance across established benchmarks (Chollet, 2017; Luo et al., 2021; Liu et al., 2021; Cao et al., 2022; Tan et al., 2024a; Yan et al., 2024a; Yang et al., 2025). Nevertheless, their performance often degrades under such distribution shifts, revealing limited generalizability beyond the domains observed during training.

These limitations have motivated recent efforts to exploit the richer vision-language understanding and reasoning capabilities of Multimodal Large Language Models (MLLMs) for deepfake detection, by strengthening their ability to identify forgery-specific evidence through strategies such as supervised fine-tuning (SFT) (Wen et al., 2025), localization-aware training (Huang et al., 2025),

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.

reinforcement-optimized reasoning (Tan et al., 2025b), and forgery-aware fine-tuning (Chen et al., 2026). These approaches provide a promising direction for generalizing across changes in fake-image distributions. However, robust generalization across both real and fake image distributions remains substantially challenging. As illustrated in Figure 7, under our Cross-domain & forgery evaluation setting, state-of-the-art MLLM-based detectors exhibit a pronounced performance asymmetry, with a substantial imbalance between real and fake recall under domain shifts. This observation suggests that improving forgery-specific discrimination alone may not be sufficient to ensure balanced and robust generalization across both sides of the detection problem, motivating the development of a deepfake detector specifically designed to handle both forms of domain shift.

Contributions. We introduce **CounterFake**, an MLLM-based deepfake detector designed to address both fake-side and real-side domain shifts through complementary fine-tuning and post-hoc stages (see Figure 1). For fine-tuning, we first construct forensic rationales using *ReForensics*, a reference-guided captioning scheme that compares fake images with their real sources or same-method fake references to identify manipulation-induced discrepancies and shared artifacts. Building on them, we propose *cross-modal counterfactual preference optimization*. With the image fixed, we alter forensic attributes in the rationale while preserving its linguistic form to construct a rejected response. With the rationale fixed, we retrieve a visually similar hard-negative image from the opposite authenticity class. The two objectives favor the original image-rationale pair over these counterfactual mismatches, providing explicit supervision on the correspondence between visual evidence and forensic descriptions. To improve robustness to unseen real-image distributions after fine-tuning, we apply *weight-space interpolation* between the pretrained and fine-tuned models. We further use *intermediate representation probing* to predict authenticity directly from the interpolated model's intermediate features, bypassing the conventional label-token readout. This regularized linear probe uses the same binary training labels and requires no further fine-tuning of the MLLM.

We conduct extensive experiments using evaluation benchmarks adopted from DF40 (Yan et al., 2024b) under two distribution-shift settings: (*i*) cross-domain generalization and (*ii*) joint cross-domain/cross-forgery generalization. For cross-domain evaluation, we evaluate across 32 forgery methods on **92K** test samples, while the joint domain/forgery setting covers 7 test sets comprising **22K** samples. Across these rigorous evaluation settings, CounterFake outperforms state-of-the-art MLLM-based deepfake detectors by **9.1** and **7.9** percentage points in average macro-F1, respectively. Notably, these gains arise from complementary improvements on the two sides of distribution-shift robustness: our counterfactual preference learning improves fake recall by **6.8** percentage points over vanilla SFT, while weight-space interpolation improves real recall by **15.4** percentage points. Together, these complementary design choices improve robustness to distribution shifts on both real and fake images, enabling stronger generalization across cross-domain and joint cross-domain/cross-forgery settings.

2 Related Work

Deepfake detection. With the advent of Vision-Language Models (VLMs), such architectures have been adopted in deepfake detection to leverage the generalization capabilities of large-scale pre-trained representations such as CLIP (Radford et al., 2021). Ojha et al. (2023) utilized CLIP, showing that linear probing on frozen image embeddings outperforms traditional methods on unseen generation methods. Building on this work, numerous approaches have leveraged this pre-trained architecture to adapt generic features for forgery classification (Baraldi et al., 2024; Koutlis & Papadopoulos, 2024; Yan et al., 2024a; Cheng et al., 2025; Jung et al., 2026). Specifically, Baraldi et al. (2024) employed contrastive learning to align global-local similarities, while Koutlis & Papadopoulos (2024) exploited intermediate backbone features. Furthermore, Effort (Yan et al., 2024a) and CoSPY (Cheng et al., 2025) focused on harnessing the rich semantic and contextual information within CLIP. Moving beyond visual-only adaptations, recent work has extended to multimodal approaches incorporating the text encoder (Khan & Dang-Nguyen, 2024; Wang et al., 2024b; Tan et al., 2025a; Zhu et al., 2026). Despite these advances, these methods remain largely discriminative, motivating explainable deepfake detection with linguistic reasoning.

To bridge this gap, recent work has increasingly leveraged MLLMs for deepfake detection. However, standard MLLMs are inherently limited in perceiving subtle, low-level manipulation traces, as their visual encoders are predominantly aligned with coarse-grained semantic concepts rather than pixel-

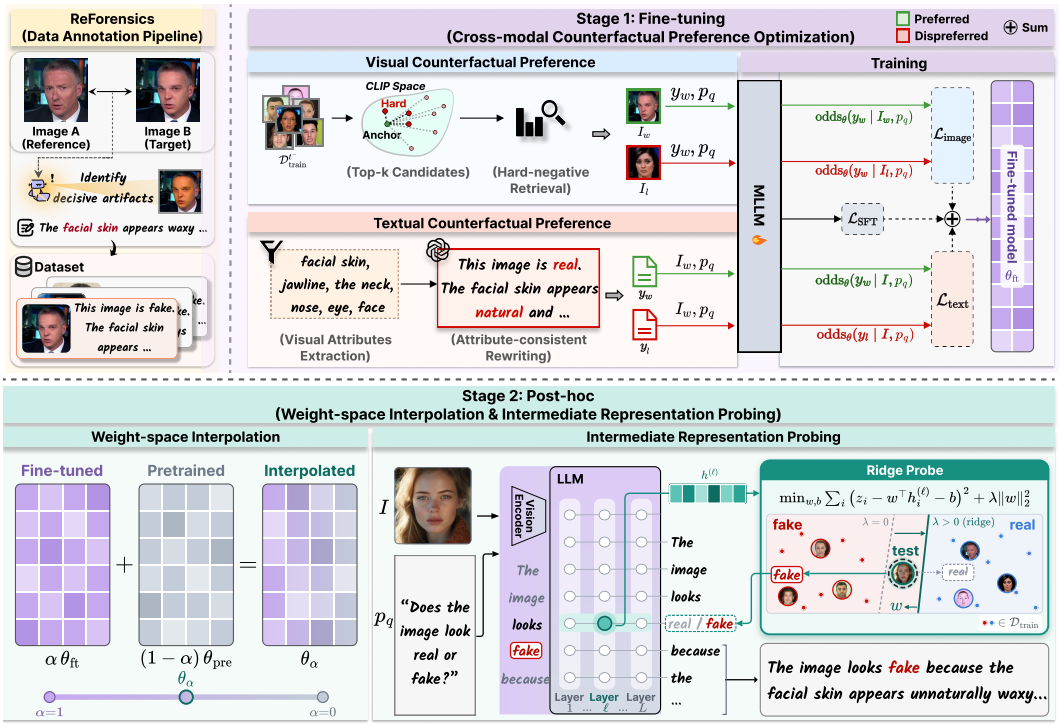


Figure 1: **Overview of CounterFake. (Upper-left)** ReForensics constructs high-quality forensic rationales through reference-guided comparison, using paired reference images and forgery-specific artifact patterns to identify manipulation-relevant visual cues. **(Upper-right)** CounterFake is trained via cross-modal counterfactual preference optimization over visual evidence and textual rationales. **(Lower)** After fine-tuning, weight-space interpolation improves robustness to unseen real-image domains, and intermediate representation probing directly reads out authenticity information from intermediate language model representations while preserving rationale generation.

level details (Tong et al., 2024; Liu et al., 2023; Li et al., 2023). In response, recent research aims to enhance the visual capability of MLLMs to capture fine-grained forgery artifacts. Specifically, Zhou et al. (2025) and Peng et al. (2025) incorporate auxiliary visual encoders to detect high-frequency details often overlooked by standard backbones. To mitigate modality interference, Zhang et al. (2024); Guo et al. (2025) independently optimize visual and text encoders before integration, while SIDA (Huang et al., 2025) uses pixel-level segmentation to focus visual attention on forgery traces. Building on these approaches, we introduce an MLLM-based deepfake detector with complementary methods for improved robustness under domain shifts.

Preference optimization for MLLM. Recent preference optimization methods align model behavior with human preferences. Many of them, such as DPO (Rafailov et al., 2023) and KTO (Ethayarajh et al., 2024), rely on a frozen reference model, whereas ORPO (Hong et al., 2024) removes the reference model by adding an odds-ratio penalty to the standard supervised fine-tuning loss. Recent works extend preference learning to MLLMs, showing that incorporating rejected images improves visual-textual alignment (Wang et al., 2024a; Luo et al., 2024; Jiang et al., 2024; Fu et al., 2025b; Wu et al., 2025; Zheng et al., 2026). Specifically, Wang et al. (2024a) construct rejected samples using cropped images, while Jiang et al. (2024) generate rejected images by adding diffusion noise to key visual objects. Wu et al. (2025) further show that semantically contradictory and visually challenging rejected images are more effective at preventing shortcut learning, encouraging MLLMs to attend to fine-grained visual details and achieve stronger visual-textual alignment.

Despite these advances, preference optimization remains underexplored in MLLM-based deepfake detection. Existing methods (Zhou et al., 2025; Tan et al., 2025b) adopt text-only preference supervision without leveraging image-level preferences. In this study, CounterFake incorporates cross-modal counterfactual preference optimization to better handle shifts across fake-image domains by contrasting counterfactual pairs at both the response and image levels.

3 METHOD

We introduce CounterFake, an MLLM-based deepfake detector designed to improve robustness through complementary fine-tuning and post-hoc procedures. We first present **ReForensics** for constructing forensic rationales (Section 3.1), then describe **cross-modal counterfactual preference optimization** (Section 3.2), followed by two post-hoc procedures, namely (i) **weight-space interpolation** and (ii) **intermediate representation probing** (Section 3.3).

3.1 REFORENSICS: REFERENCE-GUIDED CAPTIONING SCHEME FOR DEEPFAKES

We introduce ReForensics (Reference-guided Forensics Annotation; see Figure 1) to construct a preferred response y_w for each training image, consisting of the real/fake label followed by a forensic rationale. Existing approaches obtain forensic rationales through human annotation (Zhang et al., 2024) or MLLM-generated captions (Wen et al., 2025). To help distinguish forgery-specific evidence from incidental image characteristics, ReForensics augments MLLM annotation with reference-image comparisons and a taxonomy of forensic artifacts. The resulting responses provide supervision for our preference optimization framework (Section 3.2).

For each fake image, an MLLM annotator is provided with the query image, reference images, and artifact patterns corresponding to its forgery type. For paired data, the corresponding real source is used as a reference to identify manipulation-induced discrepancies between fake and real images. When paired sources are unavailable, same-method fake samples are used as references to help identify shared forgery artifacts, motivated by prior findings on generator-specific traces (Wang et al., 2020; Corvi et al., 2023a;b; Cozzolino et al., 2024; Pellegrini et al., 2025).

To guide annotation toward type-specific forensic cues, we constructed a taxonomy of 59 artifact patterns based on prior deepfake forensic studies (Pei et al., 2024; Rehaan et al., 2024), as summarized in Table 4 of Appendix B. The annotator uses the relevant patterns to guide its inspection and is instructed to describe only artifacts supported by the visual comparison, avoiding incidental artifacts and image-specific noise. For real images, the annotator instead generates a rationale describing the absence of visually supported forgery artifacts. Additional annotation details are provided in Appendix B, while human evaluation of caption quality is provided in Appendix G.

3.2 FINE-TUNING VIA CROSS-MODAL COUNTERFACTUAL PREFERENCE OPTIMIZATION

In standard preference optimization, training data are structured as triplets (x, y_w, y_l), where y_w and y_l are preferred and rejected responses for input x, respectively, with the objective of assigning a higher preference to y_w. We first question how to construct a preference dataset for deepfake detection MLLM training out of a standard dataset $\mathcal{D} = \{(x_i, t_i)\}_{i=1}^{N}$. For each training sample $(x, t) = ((I, p_q), t)$, we use the preferred response y_w constructed as described in Section 3.1.

High-quality negatives can provide fine-grained signals beyond the binary real/fake label. To this end, we construct counterfactual negatives in both modalities: *attribute-consistent rewriting* modifies the response while keeping the image fixed, whereas *hard-negative image retrieval* changes the image while keeping the preferred response fixed. These complementary constructions encourage the model to distinguish subtle inconsistencies between visual evidence and textual reasoning.

Attribute-consistent rewriting. Prior work on VLM preference optimization shows that contrasting responses without grounding the preference in visual evidence often leads models to rely on superficial linguistic cues (*e.g.*, wording or style) rather than visually grounded reasoning (Wu et al., 2025). Accordingly, we propose a two-stage LLM-based procedure that produces challenging rejected responses by selectively modifying deepfake-related visual attributes while preserving linguistic form. Given an image-prompt pair (I, p_q) and its preferred rationale y_w, we first extract a set of key semantic attributes via an LLM $\mathcal{E}_{\text{LLM}}(\cdot)$: $\mathcal{A}(y_w) = \mathcal{E}_{\text{LLM}}(y_w) = \{a_1, a_2, \ldots, a_m\}$, where each a_i describes a localized visual cue indicative of authenticity or manipulation.

Let t^- denote the class label opposite to the ground-truth label t. While preserving the sentence structure, writing style, and non-attribute content of y_w, we modify the extracted attributes and the explicit real/fake prediction to support t^-, *i.e.*, $\tilde{y}_{\text{attr}} = \mathcal{G}_{\text{LLM}}\big(y_w \mid \mathcal{A}(y_w), t^-\big)$, where $\mathcal{G}_{\text{LLM}}(\cdot)$ denotes an LLM-based generator conditioned on the counterfactual target label. We set $y_l = \tilde{y}_{\text{attr}}$,

yielding a linguistically similar counterfactual response in which both the label and the supporting forensic attributes are inconsistent with the visual evidence. The rewritten responses are further checked for quality control; see Appendix B.5.

Hard-negative image retrieval. For image-level preference learning, we retrieve visually similar images from the opposite ground-truth class. Given an anchor image I_a with prompt p_q, rationale y_w, and label t, let $\mathcal{D}^{t^-}_{\text{train}}$ denote the training images with the opposite label. Using the CLIP image encoder $f_{\text{img}}(\cdot)$, we retrieve the top-k visually similar candidates:

$$\mathcal{C}_k(I_a) = \text{Top-}k\Big(\cos\big(f_{\text{img}}(I_a), f_{\text{img}}(I)\big) \;\Big|\; I \in \mathcal{D}^{t^-}_{\text{train}} \Big), \tag{1}$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity. We use the CLIP image encoder adopted by our LLaVA-1.5 backbone.[1] Given the candidate set, we use MLLM to select the final hard negative:

$$\tilde{I}_{\text{hn}} = \underset{I \in \mathcal{C}_k(I_a)}{\text{argmax}} \; \log p_{\theta_{\text{pre}}}(y_w \mid I, p_q), \tag{2}$$

where $p_{\theta_{\text{pre}}}$ denotes the pretrained MLLM before preference optimization. Mining is performed once before preference optimization using this fixed model and is restricted to the training split. We then set $I_w = I_a$ and $I_l = \tilde{I}_{\text{hn}}$. Given these textual and visual counterfactual pairs, we formulate cross-modal counterfactual preference optimization to favor visually supported image-response pairs over their counterfactual alternatives.

Odds ratio-based preference losses. We adopt ORPO (Hong et al., 2024) to define our preference losses, although our counterfactual preference construction can also be applied to other preference optimization methods, *e.g.*, DPO (Rafailov et al., 2023). Given an input x and a response y, with $p_\theta(y \mid x)$ computed as the exponential of the mean log-probability over the label and rationale tokens, the odds-based score is defined as

$$\text{odds}_\theta(y \mid x) = \frac{p_\theta(y \mid x)}{1 - p_\theta(y \mid x)}. \tag{3}$$

The textual loss contrasts y_w with y_l for the same image, while the visual loss contrasts I_w with I_l for the same response:

$$\mathcal{L}_{\text{text}} = -\log \sigma\left(\log \frac{\text{odds}_\theta(y_w \mid I, p_q)}{\text{odds}_\theta(y_l \mid I, p_q)}\right), \quad \mathcal{L}_{\text{image}} = -\log \sigma\left(\log \frac{\text{odds}_\theta(y_w \mid I_w, p_q)}{\text{odds}_\theta(y_w \mid I_l, p_q)}\right). \tag{4}$$

Both losses therefore increase the relative odds of the visually consistent image-response pair over its counterfactual counterpart.

Overall objective. Combining the two preference losses with standard supervised fine-tuning (SFT), following ORPO, the final training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{SFT}}(y_w \mid I, p_q) + \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{image}}\mathcal{L}_{\text{image}}, \tag{5}$$

where λ_{text} and λ_{image} are hyperparameters. In our experiments, we fix $\lambda_{\text{text}} = \lambda_{\text{image}} = 0.5$.

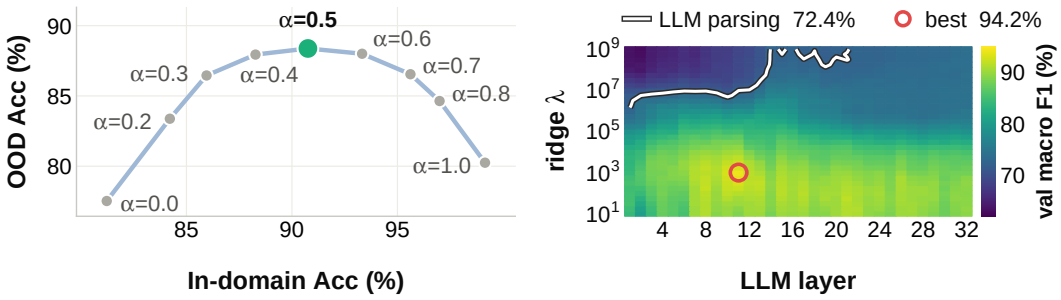
3.3 Post-hoc Robustness via Weight Interpolation & Representation Probing

We introduce two simple yet effective post-hoc procedures to improve robustness under domain shifts without additional training data or further fine-tuning of the MLLM.

Weight-space interpolation. Wortsman et al. (2022) show that interpolating pretrained and fine-tuned vision-model weights can improve robustness after fine-tuning, while Ojha et al. (2023); Yan et al. (2024a) show that preserving pretrained vision representations improves generalization to unseen forgeries in deepfake detection. Motivated by these findings, we introduce weight-space interpolation by linearly interpolating the pretrained model θ_{pre} and the fine-tuned model θ_{ft} as:

$$\theta_\alpha = (1-\alpha)\theta_{\text{pre}} + \alpha\theta_{\text{ft}}, \tag{6}$$

[1] We also test alternative retrieval approaches in Appendix F.4.



(a) (b)

Figure 2: **Weight interpolation and the token-level readout bottleneck.** (a) In-domain and out-of-distribution (cross-domain/forgery) accuracies along the weight interpolation path from the pre-trained ($\alpha = 0$) to the fine-tuned ($\alpha = 1$) weights. (b) Validation macro-F1 of a linear probe over LLM layers and ridge strengths λ. The white line indicates the token-level readout performance, while the red circle marks the best-performing configuration on the validation set (see Appendix C.1 for details). The validation-selected configuration uses layer 11 with $\lambda = 10^3$ for all experiments.

where $\alpha \in [0, 1]$ controls the interpolation ratio. As shown in Figure 2a, weight-space interpolation improves performance under the Cross-domain & forgery settings over a wide range of α. Following Wortsman et al. (2022), we simply fix $\alpha = 0.5$ for all experiments.

Intermediate representation probing. MLLM-based detectors typically predict an authenticity label as part of a structured textual response, such as *"The image looks [real/fake] because ⋯"*, followed by a rationale describing the visual evidence (Zhang et al., 2024; Wen et al., 2025). Given an image I, question prompt p_q, and fixed prefix y_{pre}, we consider the ℓ-th hidden representation $h^{(\ell)}_{\text{label}} \in \mathbb{R}^d$ at the position immediately after y_{pre}, where the authenticity label is predicted. As shown in Figure 2b, intermediate-layer representations yield better authenticity prediction than the conventional LM token-prediction pathway. Motivated by this, we introduce a probe on $h^{(\ell)}_{\text{label}}$ that bypasses the remaining LM layers and LM head to directly predict the authenticity label token.

Specifically, we consider a ridge-regularized linear probe on h_{label} after standardization:

$$(w^*, b^*) = \underset{w,b}{\operatorname{argmin}} \sum_{i=1}^{N} \left(w^\top \tilde{h}_i + b - z_i \right)^2 + \lambda \|w\|_2^2, \tag{7}$$

where $\tilde{h}_i$ is the standardized representation of h_{label} using feature-wise statistics computed from training samples, and $z_i \in \{-1, +1\}$ is the probe target derived from the binary authenticity label, with $+1$ denoting *fake*. The resulting probe defines the linear decision boundary $w^{*\top} \tilde{h} + b^* = 0$.

At inference, the predicted authenticity label from the probe is inserted at the label position, after which autoregressive generation continues to produce the subsequent rationale. This procedure requires no further fine-tuning of the MLLM and no additional training image.

4 Experiments

We conduct an extensive evaluation of CounterFake against specialized and general-purpose MLLM-based detectors.[2] We train CounterFake on 44K real and fake images from DF40 (Yan et al., 2024b), covering 30 forgery methods and annotated with our ReForensics pipeline. We evaluate generalization under two DF40 settings: *Cross-domain*, with an unseen real-source domain, and *Cross-domain & forgery*, with both unseen source domains and forgery methods. For cross-domain evaluation, we use the Celeb-DF-domain test set in DF40, comprising 92K samples spanning all 32 forgery methods, rather than a manually selected subset. The Cross-domain & forgery setting

[2] Additional results on inference efficiency, alternative retrieval schemes, backbone generalization, and corruption robustness are provided in Appendix F.

Table 1: **Comparison of accuracy and macro-F1 on Cross-domain benchmarks.** We evaluate 32 forgery methods on Celeb-DF (Li et al., 2020), which provides an unseen real-source domain distinct from those used for training. We group these methods into four forgery categories (FS, FR, EFS, and FE) and report accuracy (Acc) and macro-F1 (mF1) averaged over the methods in each category. The best and second-best results are **bolded** and underlined, respectively. Per-method accuracy and macro-F1 are reported in Appendix D.

Method	FS (Face Swapping)		FR (Face Reenactment)		EFS (Entire Face Synthesis)		FE (Facial Editing)		Avg.†	
	Acc	mF1	Acc	mF1	Acc	mF1	Acc	mF1	Acc	mF1
Generic MLLM										
LLaVA1.5-7B (Liu et al., 2024)	0.5114	0.3576	0.5025	0.3389	0.5378	0.4087	0.5036	0.3413	0.5161	0.3661
Qwen2.5-VL-7B (Bai et al., 2025)	0.5016	0.3380	0.5003	0.3351	0.5189	0.3718	0.5014	0.3373	0.5065	0.3475
InternVL3-8B (Zhu et al., 2025)	0.5469	0.4685	0.5244	0.4347	0.6128	0.5577	0.4828	0.3590	0.5571	0.4803
GPT-4o (Hurst et al., 2024)	0.5463	0.4346	0.5453	0.4381	0.5501	0.4364	0.5507	0.4371	0.5472	0.4366
MLLM-based detector										
M2F2Det (7B) (Guo et al., 2025)	0.5672	0.4547	0.5059	0.3519	0.6138	0.5355	0.6424	0.5923	0.5611	0.4457
Deep-VRM (8B) (Lin et al., 2026)	0.5150	0.5107	0.5542	0.5459	0.6691	0.6460	0.7075	0.6826	0.5839	0.5715
TranX-Adapter (7B) (Wang et al., 2026)	0.5081	0.3558	0.4991	0.3363	0.7682	0.7202	0.5793	0.4917	0.5883	0.4666
SIDA-13B (Huang et al., 2025)	0.5497	0.4285	0.5052	0.3543	0.6222	0.5446	**0.9556**	**0.9555**	0.5683	0.4534
X^2-DFD (7B) (Chen et al., 2026)	0.6936	0.6733	0.5983	0.5879	0.7169	0.6947	0.7200	0.6975	0.6660	0.6487
FakeVLM (7B) (Wen et al., 2025)	0.7705	0.7687	0.6081	0.5991	0.8003	0.7970	0.8366	0.8322	0.7210	0.7159
VERITAS (8B) (Tan et al., 2025b)	0.4468	0.3506	0.4203	0.3356	0.4366	0.3439	0.4096	0.3257	0.4325	0.3421
VERITAS++ (8B) (Tan et al., 2026)	0.5545	0.4806	0.5794	0.5213	0.6263	0.5791	0.5465	0.4756	0.5860	0.5265
CounterFake (7B)	**0.7954**	**0.7932**	**0.7767**	**0.7689**	**0.8573**	**0.8561**	0.8891	0.8878	**0.8107**	**0.8067**

† Avg. is the mean over all 32 forgery methods, not the four categories. Unparsed outputs are treated as incorrect predictions.

Table 2: **Comparison of macro-F1 on Cross-domain & forgery benchmarks.** We evaluate unseen forgery techniques across unseen real-source domains. The best and second-best results are **bolded** and underlined, respectively. We also report comparison in accuracy in Appendix D.

Method	DeepFaceLab	HeyGen	WhichIsReal	CollabDiff	StarGAN	StarGANv2	StyleCLIP	Avg.
Generic MLLM								
LLaVA1.5-7B (Liu et al., 2024)	0.3078	0.3344	0.3336	0.3385	0.5205	0.3436	0.3179	0.3566
Qwen2.5-VL-7B (Bai et al., 2025)	0.3093	0.3652	0.3350	0.3355	0.4296	0.3408	0.3032	0.3455
InternVL3-8B (Zhu et al., 2025)	0.5420	0.7088	0.3926	0.3431	0.6111	0.3580	0.3390	0.4707
GPT-4o (Hurst et al., 2024)	0.3960	0.6577	0.7614	0.7208	0.4379	0.5372	0.7675	0.6112
MLLM-based detector								
M2F2Det (7B) (Guo et al., 2025)	0.4388	0.3619	0.3387	0.3510	0.7608	0.3800	0.4078	0.4341
Deep-VRM (8B) (Lin et al., 2026)	0.5180	0.7437	0.7430	0.6646	0.4758	0.3980	**0.9807**	0.6463
TranX-Adapter (7B) (Wang et al., 2026)	0.2978	0.3048	0.6543	**0.9810**	0.5723	0.3792	0.3754	0.5093
SIDA-13B (Huang et al., 2025)	0.3619	0.6008	0.3530	0.7453	0.7536	0.5879	0.9065	0.6156
X^2-DFD (7B) (Chen et al., 2026)	0.6851	0.5805	0.3783	0.3616	0.4408	0.4023	0.4260	0.4678
FakeVLM (7B) (Wen et al., 2025)	0.7867	0.6722	0.7939	0.8850	0.8862	0.6376	0.9582	0.8028
VERITAS (8B) (Tan et al., 2025b)	0.4211	0.7687	**0.9303**	0.8484	0.6458	0.4827	0.5449	0.6631
VERITAS++ (8B) (Tan et al., 2026)	0.5009	0.7420	0.6082	0.6558	0.7461	0.4231	0.7275	0.6291
CounterFake (7B)	**0.8816**	**0.8056**	0.9010	0.9028	**0.9104**	**0.8489**	0.9191	**0.8813**

covers seven of the remaining eight forgery methods across four unseen real-source domains, with MidJourney excluded from evaluation.[3] See Appendix C for full experimental details.

4.1 Results

Comparative results. As shown in Tables 1 and 2, CounterFake achieves the best average performance in both settings. In the Cross-domain setting, it achieves an accuracy of 0.8107 and macro-F1 of 0.8067, outperforming FakeVLM by 8.97 and 9.08 percentage points, respectively. It ranks first on both metrics for FS, FR, and EFS and second on FE, with its lowest category-level macro-F1 remaining at 0.7689. In the Cross-domain & forgery setting, CounterFake achieves the best average macro-F1 of 0.8813, surpassing FakeVLM by 7.85 percentage points, and ranks in the top two on six of seven forgery methods. Notably, while FakeVLM drops from 0.8862 on StarGAN to 0.6376 on StarGANv2, CounterFake achieves 0.9104 and 0.8489, respectively, demonstrating more consistent performance across unseen forgeries. Despite sharing the LLaVA-1.5-7B backbone with FakeVLM, CounterFake achieves substantially better generalization while training only approximately 5% of the parameters. A controlled comparison with five vision-only detectors retrained on the same training set is provided in Appendix E.

[3] MidJourney is included in the training data used by the FakeVLM and VERITAS baselines. To ensure a consistent evaluation protocol, we therefore also include it in our training set.

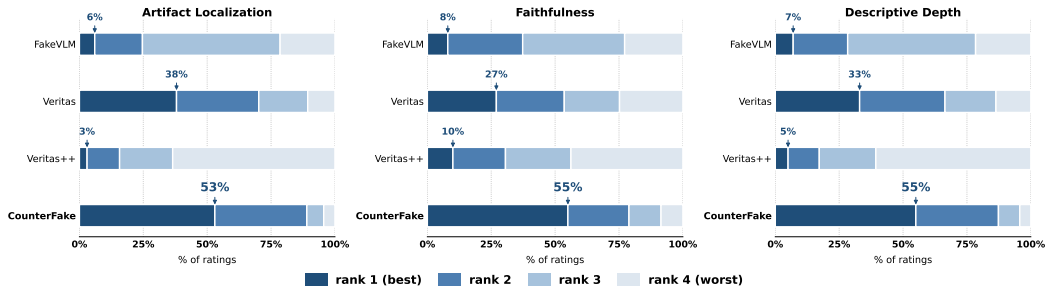


Figure 3: **Human evaluation of model-generated rationales.** We compare the quality of model-generated forensic rationales through a human preference study. Participants rank four anonymized rationales (CounterFake, VERITAS, VERITAS++, and FakeVLM, presented in random order) from rank-1 (best) to rank-4 (worst) for each forged image, based on artifact localization, faithfulness, and descriptive depth.

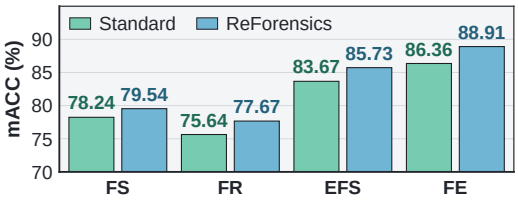


Figure 4: **Performance of ReForensics rationales.** ReForensics offers effective forensic rationale supervision in the cross-domain setting.

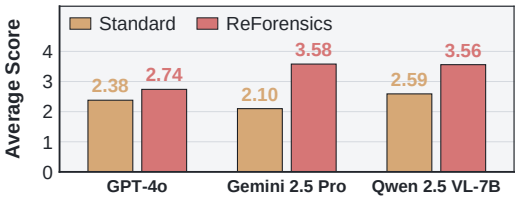


Figure 5: **Reliability of ReForensics rationales.** ReForensics receives higher caption reliability scores from the evaluation MLLMs.

Human evaluation. To further assess the quality of the generated forensic rationales beyond detection accuracy, we conduct a human evaluation with 32 participants on 20 fake images that are correctly classified by all compared models. By evaluating only samples that are correctly classified by all compared models, we focus the comparison on the quality of the generated forensic rationales rather than label correctness. Participants evaluate the rationales in terms of artifact localization, faithfulness, and descriptive depth. As shown in Figure 3, CounterFake achieves the highest proportion of rank-1 ratings across all three criteria. In particular, CounterFake consistently receives higher rankings than the other models across artifact localization, faithfulness, and descriptive depth. Together with the qualitative analysis in Figure 6, these results indicate that CounterFake improves not only detection performance but also the quality of model-generated forensic rationales. Full evaluation details are provided in Appendix G.

ReForensics analysis. We compare ReForensics with a "no-reference" captioning baseline, following the single-image prompt design of FakeVLM (Wen et al., 2025). Compared with standard captioning, ReForensics incorporates reference-guided comparison and forensic expertise prompts derived from our artifact taxonomy, encouraging captions to focus on decisive and shared forgery cues. Figure 4 shows that ReForensics rationales lead to stronger performance than standard captions across FS, FR, EFS, and FE, demonstrating their effectiveness as forensic rationale supervision for CounterFake. We further evaluate caption reliability through multiple evaluation MLLMs on 500 random samples from DF40 (Yan et al., 2024b), using the prompt in Figure 16 of Appendix G. As shown in Figure 5, ReForensics receives higher reliability scores than standard captions, indicating more reliable forensic rationales. We also conduct a human evaluation of caption quality, with further analysis in Appendix G.

Qualitative analysis. Figure 6 compares the rationales generated by CounterFake, FakeVLM, VERITAS, and the base LLaVA-1.5-7B. LLaVA-1.5-7B produces nearly identical responses for both examples, relying on generic cues such as lighting consistency, facial proportions, and boundary artifacts. FakeVLM considers more localized facial details, but still misclassifies both images as real and fails to recognize the actual manipulation cues. VERITAS provides more forensic-oriented explanations, yet some of its cues are weakly grounded, such as a skin-tone mismatch around the eyes or distortions in the nose and mouth. In contrast, CounterFake accurately localizes subtle

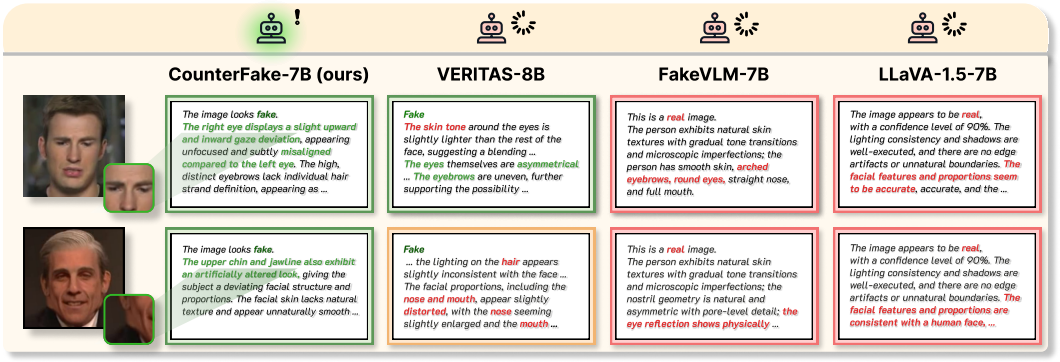


Figure 6: **Qualitative comparison.** CounterFake produces more precise and visually grounded rationales that explicitly identify deepfake artifacts. More results are provided in Appendix H.

Table 3: **Ablation studies.** Evaluated on the Cross-domain & forgery setting.

(a) Training objectives

SFT	$\mathcal{L}_{\text{text}}$	$\mathcal{L}_{\text{image}}$	Acc	mF1
✓	✓	✓	**0.8843**	**0.8813**
✓	✗	✓	0.8660	0.8606
✓	✓	✗	0.8636	0.8588
✓	✗	✗	0.8479	0.8439
✗	✗	✗	0.7614	0.7455

(b) Model components

Method	Acc	mF1
CounterFake (Ours)	**0.8843**	**0.8813**
w/o weight interp.	0.8209	0.8110
w/o interm. probing	0.8025	0.7927

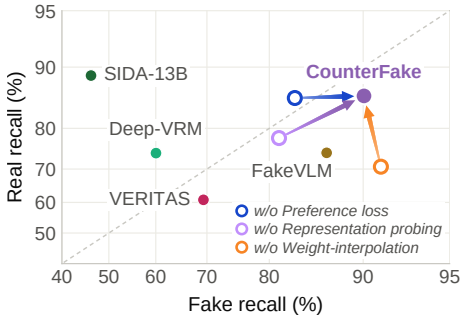


Figure 7: **Real-fake recall balance under the Cross-domain & forgery setting.** Each component contributes complementary gains to real and fake recalls.

artifacts, including gaze misalignment and degraded eyebrow details in the first example, and an altered jawline and unnaturally smooth skin in the second. These results show that CounterFake produces more visually grounded rationales. More results are provided in Appendix H.

Ablation study. We perform a component-wise analysis to assess the contributions of CounterFake components.[4] As shown in Table 3a, each preference objective improves over the SFT baseline, and combining both yields the best performance. Table 3b shows that weight interpolation and intermediate probing are both essential, with their removal reducing accuracy by 6.3%p and 8.2%p, respectively. Figure 7 highlights their complementary effects, showing distinct contributions to robustness under shifts in real and fake domains. Taken together, these results show that the proposed components jointly mitigate the asymmetric degradation in real and fake recall observed in baseline detectors, enabling CounterFake to maintain strong, balanced performance under domain shifts.

5 Conclusion

This paper introduces CounterFake, an MLLM-based deepfake detector for robust generalization under domain and forgery shifts. It combines ReForensics-based forensic supervision and cross-modal counterfactual preference optimization with two post-hoc strategies, weight-space interpolation and intermediate representation probing, to address distribution shifts on both real and fake images. Across cross-domain and joint cross-domain/forgery evaluations, CounterFake outperforms existing MLLM-based detectors by 9.1 and 7.9 percentage points in average macro-F1, respectively. Ablation results further show that the fine-tuning and post-hoc components provide complementary gains in robustness. Together, these results demonstrate the effectiveness of CounterFake in improving the robustness and generalization of MLLM-based deepfake detection under distribution shifts.

[4] Additional analyses, including feature space analysis, effect of hard-negative pool size, training stability, and comparison with other multimodal preference optimization methods are provided in Appendix F.

AI USE STATEMENT

We used Gemini 2.5 Flash to generate fake-image forensic rationales with ReForensics, and GPT-4o-mini to generate real-image rationales and counterfactual rewrites, as described in Sections 3.1 and 3.2 and Appendix B. We also used AI assistance for parts of the experimental implementation and debugging, as well as for polishing the wording and clarity of the manuscript text.

ETHICS STATEMENT

This work is intended for defensive applications of deepfake detection, including digital forensics, misinformation detection, and media authentication. As discussed in Appendix A, existing deepfake datasets may have limited demographic and environmental diversity, and forensic insights could potentially be misused to improve forgery generation. Our work focuses on detection and explanation rather than synthetic media generation.

REPRODUCIBILITY STATEMENT

Section 3 describes ReForensics, cross-modal counterfactual preference optimization, weight-space interpolation, and intermediate representation probing. Appendix B provides details on data construction and annotation, while Table 5 summarizes the implementation, training, and post-hoc inference settings. Evaluation protocols and baseline configurations are provided in Appendix C, with additional ablations and robustness analyses in Appendix F.

REFERENCES

Ravi Kumar Amaresam. Deepfake detection and ai's role in preventing digital fraud. *International Journal of Research and Applied Innovations*, 7(4):11096–11107, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Lorenzo Baraldi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara. Contrasting deepfakes diffusion via contrastive learning and global-local similarities. In *European Conference on Computer Vision*, pp. 199–216. Springer, 2024.

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.

Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18689–18698. IEEE, 2022.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

Yize Chen, Zhiyuan Yan, Guangliang Cheng, Kangran Zhao, Siwei Lyu, and Baoyuan Wu. X2-dfd: A framework for explainable and extendable deepfake detection. *Advances in Neural Information Processing Systems*, 38:83792–83839, 2026.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13455–13465, 2025.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.

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.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: From generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 973–982, June 2023a.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4356–4366, 2024.

Cassandra Cross. Using artificial intelligence (ai) and deepfakes to deceive victims: the need to rethink current romance fraud prevention messaging. *Crime Prevention and Community Safety*, 24(1):30–41, 2022.

Audrey de Rancourt-Raymond and Nadia Smaili. The unethical use of deepfakes. *Journal of Financial Crime*, 30(4):1066–1077, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Jinlan Fu, Shenzhen Huangfu, Hao Fei, Yichong Huang, Xiaoyu Shen, Xipeng Qiu, and See-Kiong Ng. Mcm-dpo: Multifaceted cross-modal direct preference optimization for alt-text generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5355–5364, 2025a.

Jinlan Fu, Shenzhen Huangfu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. *arXiv preprint arXiv:2501.16629*, 2025b.

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 Computer Vision and Pattern Recognition Conference*, pp. 105–116, 2025.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

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 Computer Vision and Pattern Recognition Conference*, pp. 28831–28841, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Songtao Jiang, Yan Zhang, Ruizhe Chen, Tianxiang Hu, Yeying Jin, Qinglin He, Yang Feng, Jian Wu, and Zuozhu Liu. Modality-fair preference optimization for trustworthy mllm alignment. *arXiv preprint arXiv:2410.15334*, 2024.

Inho Jung, Hyeongjun Choi, Binh M Le, Hohyun Na, and Simon S Woo. A rich knowledge space for scalable deepfake detection. In *The Fourteenth International Conference on Learning Representations*, 2026.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 1006–1015, 2024.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, pp. 394–411. Springer, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3207–3216, 2020.

Kaiqing Lin, Zhiyuan Yan, Ruoxin Chen, Ke-Yue Zhang, Yue Zhou, Caiyong Piao, Bin Li, Taiping Yao, Bo Wang, Youchang Xiao, et al. Deep residual injection for full-spectrum forensic signal perception in multimodal large language models. *arXiv preprint arXiv:2606.15880*, 2026.

Leo SF Lin. Examining the role of deepfake technology in organized fraud: Legal, security, and governance challenges. *Frontiers in Law*, 4:6–17, 2025.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2(3):12, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 772–781, 2021.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in vlms. *arXiv preprint arXiv:2501.00569*, 2024.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16317–16326, 2021.

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.

Jinhee Park, Guisik Kim, Choongsang Cho, and Junseok Kwon. Real-aware residual model merging for deepfake detection. *arXiv preprint arXiv:2509.24367*, 2025.

Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *ACM Computing Surveys*, 2024.

Lorenzo Pellegrini, Davide Cozzolino, Serafino Pandolfini, Davide Maltoni, Matteo Ferrara, Luisa Verdoliva, Marco Prati, and Marco Ramilli. Ai-genbench: A new ongoing benchmark for ai-generated image detection. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. IEEE, 2025.

Siran Peng, Zipei Wang, Li Gao, Xiangyu Zhu, Tianshuo Zhang, Ajian Liu, Haoyuan Zhang, and Zhen Lei. Mllm-enhanced face forgery detection: A vision-language fusion solution. *arXiv preprint arXiv:2505.02013*, 2025.

Viacheslav Pirogov. Visual language models as zero-shot deepfake detectors. *arXiv preprint arXiv:2507.22469*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Qi Qin, Le Zhuo, Yi Xin, Ruoyi Du, Zhen Li, Bin Fu, Yiting Lu, Xinyue Li, Dongyang Liu, Xiangyang Zhu, et al. Lumina-image 2.0: A unified and efficient image generative framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20031–20042, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Mansi Rehaan, Nirmal Kaur, and Staffy Kingra. Face manipulated deepfake generation and recognition approaches: A survey. *Smart Science*, 12(1):53–73, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Akshay Shetye, Nilakshi Jain, Shwetambari Borade, and Vineet Kumar. Deepfake technologies in financial fraud: Generation, detection, and mitigation strategies. In *2025 Global Conference in Emerging Technology (GINOTECH)*, pp. 1–6. IEEE, 2025.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5052–5060, 2024a.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024b.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025a.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Chuanbiao Song, Senyuan Shi, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas: Generalizable deepfake detection via pattern-aware reasoning. *arXiv preprint arXiv:2508.21048*, 2025b.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Zijian Yu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas++: Value-aware on-policy distillation for perception-enhanced aigi detection. *arXiv preprint arXiv:2607.27113*, 2026.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, 2024a.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Wenbin Wang, Yuge Huang, Jianqing Xu, Yue Yu, Jiangtao Yan, Shouhong Ding, Pan Zhou, and Yong Luo. Tranx-adapter: Bridging artifacts and semantics within mllms for robust ai-generated image detection. *arXiv preprint arXiv:2602.21716*, 2026.

Yabin Wang, Zhiwu Huang, Su Zhou, Adam Prugel-Bennett, and Xiaopeng Hong. Penny-wise and pound-foolish in deepfake detection. *arXiv preprint arXiv:2408.08412*, 2024b.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4129–4138. IEEE, 2023.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *arXiv preprint arXiv:2503.14905*, 2025.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7949–7961. IEEE, 2022.

Shengguang Wu, Fan-Yun Sun, Kaiyue Wen, and Nick Haber. Symmetrical visual contrastive optimization: Aligning vision-language models with minimal contrastive images. *arXiv preprint arXiv:2502.13928*, 2025.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

Zhiyuan Yan, Jiangming Wang, Zhendong Wang, Peng Jin, Ke-Yue Zhang, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024a.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, 37:29387–29434, 2024b.

Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. Dˆ 3: Scaling up deepfake detection by learning from discrepancy. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23850–23859, 2025.

Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *European conference on computer vision*, pp. 399–415. Springer, 2024.

Zhixiao Zheng, Zheren Fu, Zhiyuan Yao, Dongming Zhang, and Zhendong Mao. Cat-po: Cross-modal adaptive token-rewards for preference optimization in truthful multimodal llms. In *International Conference on Learning Representations*, volume 2026, pp. 113687–113702, 2026.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. *arXiv preprint arXiv:2507.02664*, 2025.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pp. 650–667. Springer, 2022.

Jiawen Zhu, Yunqi Miao, Xueyi Zhang, Jiankang Deng, and Guansong Pang. Unleashing vision-language semantics for deepfake video detection. *arXiv preprint arXiv:2603.24454*, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Limitations and Broader Impact

Limitations. Although our experiments cover diverse forgery types and cross-domain settings, the source data is still partly shaped by existing deepfake datasets such as FaceForensics++ (Rossler et al., 2019). While such datasets are widely used in deepfake detection research, they may not fully reflect the demographic and environmental diversity of real-world scenarios. Future work could further expand the research scope by incorporating larger-scale datasets with more balanced coverage across demographic attributes such as ethnicity, age, and gender, as well as lighting conditions, camera quality, and acquisition environments. This would enable a more comprehensive evaluation of fairness, robustness, and generalization in practical deepfake detection settings.

In addition, ReForensics introduces reference-guided comparative annotation, which assumes the availability of suitable reference images. For wild fake images whose real source or generation process is unavailable, obtaining reliable source-real references or reference samples generated by the same generator can be challenging. One practical remedy is to generate fake counterparts from available real wild images using known manipulation or generation methods, and then apply ReForensics to construct reference-guided forensic rationales.

Broader impact. This work aims to advance multimodal learning for explainable deepfake detection by improving visual-textual alignment and strengthening the model's ability to identify subtle forensic artifacts. By providing both detection decisions and forensic rationales, our approach can support digital forensics, misinformation detection, and media authentication, thereby contributing to increased trust in online visual content. At the same time, as with many advances in forgery detection, insights into forensic cues could potentially be misused to inform more sophisticated forgery generation. However, our work is primarily intended for defensive purposes, and the proposed method focuses on robust detection and explanation rather than synthetic media generation. Moreover, the forensic rationales produced by our framework are grounded in comparative reasoning over subtle forensic artifacts, rather than exposing simple rules that can be directly exploited. We believe that the societal benefits of improving robustness, generalization, and transparency in deepfake detection outweigh the potential risks.

B Training Data Construction and ReForensics Annotation

B.1 Training Data Construction

In this section, we provide a detailed description of the dataset construction process employed in this study. While recent MLLM-based deepfake detection studies widely utilize the DD-VQA dataset (Zhang et al., 2024), which relies on human annotations, it inherently suffers from limited linguistic diversity and scalability due to the prohibitive cost of manual annotation. To overcome these limitations and construct a larger-scale, more diverse training corpus, we build our data primarily on DF40 (Yan et al., 2024b). Specifically, we use the 29 forgery methods included in the DF40 training split and additionally include MidJourney, resulting in a total of 30 forgery methods for training. For validation, we use fake samples generated by three held-out forgery methods, MCNet, SadTalker, and Wav2Lip, on the FaceForensics++ (Rossler et al., 2019) source domain. These three methods are excluded from training and used only for validation.

The final training set contains 44,744 images. Fake-image captions are generated with Gemini 2.5 Flash using ReForensics, while real-image captions and counterfactual rewrites use GPT-4o-mini. The captions are manually inspected and filtered, and the final set is resampled to retain class balance. We use the separate validation set described in Appendix C.1 for model selection. MidJourney is included in training and excluded from evaluation to follow the training-data convention used by the FakeVLM and VERITAS baselines, as explained in the main paper.

B.2 Reference Selection Strategy for ReForensics

To obtain high-quality forensic captions corresponding to the diverse manipulation techniques in DF40, we utilized our proposed **ReForensics** pipeline. This reference-guided approach ensures objective annotation by anchoring descriptions to comparative evidence. The annotation process is categorized into two strategies based on the availability of a direct reference source:

- **Strategy 1: Paired real reference (Figure 8).** For deepfake samples where the corresponding original source is available (e.g., face swapping or reenactment), we utilized the exact original real image as a paired reference. As illustrated in the prompt template, explicitly comparing the fake image with its ground-truth source allows the annotator to objectively identify manipulation artifacts, thereby minimizing subjective hallucinations.
- **Strategy 2: Paired fake reference (Figure 9).** For fully synthesized images (e.g., diffusion-based generation) where a direct pixel-wise real reference is unavailable or ill-defined, we adopted a counterpart comparison strategy. We utilized another deepfake sample generated by the identical model/method as the reference, as shown in Figure 9. By comparing two distinct samples from the same generator, the model can isolate and extract "shared artifacts" or common fingerprints inherent to that specific generative mechanism.

The base structure of the prompts used in this pipeline was adapted and refined from the methodology proposed in FakeVLM (Wen et al., 2025). Specifically, for authentic (real) images where no comparative reference is applicable, we adhered to their original protocol. By augmenting this foundation with our reference-guided paradigm for manipulated data, we facilitated the generation of more reliable and objectively grounded forensic descriptions.

B.3 Forgery Artifact Taxonomy

Table 4 presents the forgery-type taxonomy used in ReForensics. It organizes 59 forensic artifact patterns across four manipulation types: Face Swapping (FS), Face Reenactment (FR), Entire Face Synthesis (EFS), and Facial Editing (FE). The taxonomy summarizes representative visual artifacts associated with each forgery type.

B.4 ReForensics Annotation Prompts

Figures 8 and 9 show the annotation prompts used in ReForensics. The prompts guide the model to compare the target image with a reference and report only clear manipulation artifacts that satisfy the predefined taxonomy.

B.5 Quality Control

For fake-image annotation, Gemini 2.5 Flash compares the image with its reference and reports at most three visually supported artifacts consistent with the taxonomy in Table 4. If none qualifies, it may output "No artifacts found." For real images, GPT-4o-mini describes only visible facial regions. These annotations form the preferred response y_w.

For counterfactual rewriting, GPT-4o-mini uses y_w to generate a rejected response y_l for the opposite authenticity label. It changes the label and corresponding attribute descriptions while preserving other details, sentence order, and style; the image remains unchanged. Real-image annotations and real-to-fake rewrites are checked for the required opening and a length of 25–60 words, chosen to approximately match the length distribution of the original annotations. Outputs that fail these checks are regenerated up to three times and then discarded. For fake-to-real rewrites, the required opening is inserted deterministically during post-processing.

Table 4: **Forgery-type taxonomy for ReForensics.** Three experts constructed artifact taxonomies for Face Swapping (FS), Face Reenactment (FR), Entire Face Synthesis (EFS), and Facial Editing (FE), guiding ReForensics toward type-specific and discriminative forgery cues.

FS: Face Swapping	FR: Face Reenactment
Blending & boundary discontinuity • Boundary irregularity • Tone & color mismatch • Halo effect **Spectrum & texture mismatch** • Facial texture smoothing • Frequency & noise discrepancy • Compression & grid artifacts **Illumination & shadow inconsistency** • Shadow inconsistency • Shadow omission **Anatomical anomalies** • Facial feature inconsistency • Facial feature misplacement • Gaze direction mismatch **Segmentation & occlusion errors** • Blended artifacts • Transparent objects	**Structural misalignment** • Neck-head inconsistency • Iris distortion • Inconsistent shadow • Attribute loss **Boundary & masking anomalies** • Artificial seam line • Double edges / ghosting **Oral anatomical anomalies** • Dental structural malformation • Cavity texture blurring **Expression disharmony** • Static upper face • Muscle activation mismatch **Spectrum & texture mismatch** • Synthetic skin texture • Blurry lip boundary **Spurious motion artifacts** • Object / background distortion • Hair texture swimming
EFS: Entire Face Synthesis	**FE: Facial Editing**
Optical physics violations • Uniform sharpness • Illogical defocus • Resolution mismatch **Material & surface texture bias** • Digital airbrushing / wax-like skin • Painterly / muddy texture • Highlight clipping **Structural asymmetry** • Mismatched accessories • Irregular iris • Gaze deviation **Anatomical anomalies** • Dental anomalies • Ear structure anomalies • Limb / hair anomalies **Contextual & spatial logic errors** • Semantic pattern failure • Clothing pattern distortion • Hair & facial hair distortion • Background structure distortion • Spatial incoherence / blobs • Surreal / abstract background • Object hallucination **Aesthetic & lighting bias** • Hyper-realistic / volumetric lighting • Perfect symmetry • Compositional perfection	**Structural & semantic inconsistency** • Object distortion • Background distortion • Anatomical disharmony **Texture & frequency degradation** • Over-smoothed texture • Noise & sharpness discrepancy **Residual & boundary artifacts** • Ghosting artifact • Texture fusion **Generative model-specific artifacts** • Material & surface texture bias • Contextual illumination mismatch • Logical hallucination

Reforensics Captioning (Original Real Source)

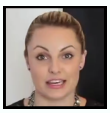


Image A
(Reference)

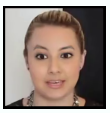


Image B
(Target)

```
# Fake
Role. You are an expert AI forensics analyst specializing in deepfake detection.

I will provide two images:
1. **Reference Image:** The **ABSOLUTE GROUND TRUTH**. This defines what is "natural" for this specific subject.
2. **Investigation Image:** The manipulated (fake) image generated using the **'{forgery_type}'** technique.

**Context:**
Face Swapping (FS)** is a technique for synthesizing the identity of a source person by transplanting it into the face area of a target image.

**[Knowledge Base: Forensic Artifact Taxonomy ({forgery_type})]**
Classify strictly using these fake artifact taxonomy.

{fake_artifact_taxonomy}

**[CRITICAL INSTRUCTION: The Differential & Quality Protocol]**
Before generating any output, you must perform the following logical steps internally:

1. **Step 1: Subtraction (Difference Check).** Compare the Investigation Image against the Reference Image. Identify what is new or different.
2. **Step 2: Validation (The "Defect" Rules).** Validate the identified differences strictly against these three mandatory rules:
   * **Rule A (Natural Trait Check):** If the feature exists in the Reference Image, it is **NOT** an artifact. **IGNORE IT.**
   * **Rule B (Visual Realism Check):** Even if a feature is NEW (not in Reference), does it look visually natural?
      * If it is a realistic change (e.g., mouth open naturally, slight head turn that looks physically correct), **IGNORE IT.** We are looking for **FLAWS**, not just
changes.
   * **Rule C (Taxonomy Compliance Check):** Can you classify this error strictly into one of the 5 categories in the Knowledge Base?
      * If the anomaly cannot be clearly defined by the provided Visual Cues (e.g., "it just looks weird" but no specific flaw), **DO NOT REPORT IT.**
      * **Do not force a classification.** If it doesn't fit the rules, discard it.
3. **Step 3: Description.** Describe ONLY the validated technical artifacts strictly as visual flaws within the Investigation Image.

**Instructions for Output:**
1. **Strict Selection:** Identify the **most distinct technical artifacts (1 to 3 maximum)** that passed ALL Steps (1, 2, and 3).
   * **Zero Tolerance for Ambiguity:** If an artifact is subtle or debatable, do not report it.
   * **Do NOT force 3 artifacts:** If only 1 distinct technical flaw exists, report only 1.
   * **Analysis Rejection Option:** If you cannot find any valid, distinct artifacts that meet the strict criteria above, output exactly: **"No artifacts found."**
2. **Format:** Use the structure below.

**[Output Structure & Example]**
*(Describe the visual flaw strictly. Do not mention "Reference Image" in the final text.)*
*(Description has only one sentence)*

**Artifact 1**
- **Where:** ...
- **Category:** ...
- **Visual Cue:** ...
- **Description:** ...
"""
```

Figure 8: **ReForensics prompt template for real-source references.** ReForensics uses the original real source image as the reference for comparative forensic annotation.

Reforensics Captioning (Same Manipulation Technique)



Image A
(Reference)



Image B
(Target)

Fake
You are an expert AI Forensics Analyst specializing in Deepfake Detection.

I will provide two images:
1. **Investigation Image A:** A synthetic face generated by a specific AI model.
2. **Investigation Image B:** Another synthetic face generated by the **SAME** AI model.

Context:
Both images were created using the **'{forgery_type}'** technique (Entire Face Synthesis).
EFS generates non-existent identities from random noise. Artifacts typically fall into two main spectrums depending on the model architecture:
* **GAN-based (e.g., StyleGAN):** Excellent skin texture (pores visible) but prone to **structural asymmetry** (mismatched earrings/eyes) and **abstract/incoherent backgrounds** (blobs).
* **Diffusion-based (e.g., Midjourney, Flux):** High aesthetic quality (cinematic lighting) but prone to **physical logic failures** (blending objects, gravity defiance) and **painterly/smoothed textures** in detailed areas.
Objective:
Perform a comparative forensic analysis to identify synthetic artifacts.
1. Identify exactly **one "Shared Artifact"** that appears in **BOTH** images, considering the {forgery_type} (AI model specific artifacts).
2. Analyze **Image A** individually to detect **up to 2 distinct artifacts** (excluding the Shared Artifact).
3. Analyze **Image B** individually to detect **up to 2 distinct artifacts** (excluding the Shared Artifact).

[Knowledge Base: Forensic Artifact Taxonomy {forgery_type}]
Classify strictly using these fake artifact taxonomy.

{fake_artifact_taxonomy}

[CRITICAL INSTRUCTION: The Analysis Protocol]
Execute the following steps internally:
1. **Step 1: The "Focus & Logic" Filter (Anti-Hallucination).**
* **Scan for Blur:** Identify areas that are out-of-focus due to depth-of-field (background, ears, shoulders, hair edges).
* **STRICT PROHIBITION:** You are **strictly prohibited** from flagging an area as *"Contextual & Spatial Logic Errors"* or *"Anatomical Anomalies"* if it is blurred or out-of-focus.
2. **Step 2: Commonality Check (Priority).** Scan both images to find an artifact category/cue that exists in **BOTH** (e.g., Wax-like Skin). Select this as the **Shared Artifact**.
3. **Step 3: Individual Selection (Anti-Redundancy).**
* Select max 2 unique artifacts for **Image A**.
* Select max 2 unique artifacts for **Image B**.
* **Constraint:** To maximize coverage, **PRIORITIZE artifacts that are DIFFERENT from the Shared Artifact.**
* *Example:* If "Wax-like Skin" is the Shared Artifact, look for other errors for the Individual Image sections. Only list "Wax-like Skin" again if it is the absolutely ONLY flaw visible.
4. **Step 4: Drafting.** Write descriptions.
* **Rule:** Even for the Shared Artifact, describe it as if observing a single instance of a flaw. Do NOT use comparison phrases like "Both images show...".

Instructions for Output:
1. **Selection Strategy:**
* **Shared Artifact:** 1 Mandatory Artifact (present in both).
* **Image A/B:** Max 2 Artifacts **(Preferably distinct from the Shared Artifact)**.
2. **Zero Tolerance:** If an artifact is subtle or debatable, do not report it.
3. **Rejection:** If no valid artifacts are found in a section, output exactly: **"No artifacts found."**
4. **Format:** Use the exact structure below. Description has only one sentence.
5. **Visual Cue:** If you can not find Visual Cue in the knowledge base, but fits Category definition, write Category name (e.g., "Material & Surface Texture Bias")

[Output Structure & Example]
[Analysis: Investigation Image A]
Artifact 1
- **Where:** ...
- **Category:** ...
- **Visual Cue:** ...
- **Description:** ...

Figure 9: ReForensics prompt template utilizing a paired deepfake generated with the same manipulation technique.

C Experimental Setup

C.1 Evaluation Datasets and Protocols

Validation set. For the vision-only detector comparison, all five baselines in Appendix E are trained using all 44,744 labeled training images. For checkpoint selection, we use a separate validation set of 3,370 images from real CelebV-HQ (Zhu et al., 2022) images and FF++ forgeries generated with MCNet, SadTalker, and Wav2Lip, which belong to the same domain as the data used for training. CelebV-HQ is absent from the training and test sets; the validation images and video directories do not overlap with the training data, and no validation image is used for testing. The three validation forgery methods are absent from training and from the seven joint cross-domain and cross-forgery test sets. They also appear by method name among the 32 cross-domain subsets, which use Celeb-DF as the real-image source.

Evaluation sets. The cross-domain evaluation covers 92K images and 32 forgery methods paired with Celeb-DF as an unseen real-source domain. The joint cross-domain and cross-forgery evaluation contains seven unseen forgery methods (DeepFaceLab, HeyGen, WhichIsReal, CollabDiff, StarGAN, StarGANv2, and StyleCLIP) and four unseen real-source domains. Accuracy and macro-F1 are reported for each test set; the cross-domain Avg. is the arithmetic mean across its 32 forgery methods, rather than across the four forgery categories.

C.2 Baselines and Evaluation Configuration

In our main comparison, we conduct an extensive evaluation against a broad range of baselines, covering both specialized and general-purpose approaches.

MLLM-based detector.

- **M2F2Det** (Guo et al., 2025) extends generalizable face forgery detection to interpretable multi-modal detection by jointly producing a fake-probability score and a textual explanation. It adapts a frozen CLIP via Forgery Prompt Learning, which optimizes universal forgery prompts and layer-wise forgery tokens to generate forged attention maps, and connects the CLIP image encoder to both a deepfake detector and an LLM through a Bridge Adapter that supplies a frequency-based token for grounded explanation generation.
- **SIDA-13B** (Huang et al., 2025) unifies social-media image deepfake detection, tampered-region localization, and textual explanation within a single multi-modal framework. Built on top of a vision-language model, it expands the vocabulary with dedicated <DET> and <SEG> tokens whose hidden states drive a detection head and a mask decoder respectively, while the model jointly produces natural-language explanations describing the basis of its judgment.
- **FakeVLM** (Wen et al., 2025) casts synthetic and deepfake image detection as an explanatory VQA task, where the model jointly issues an authenticity verdict and a natural-language artifact description. Fine-tuned end-to-end on LLaVA-v1.5 with category-aware artifact captions, it attains expert-level classification accuracy without relying on any auxiliary detector or classifier head.
- **VERITAS** (Tan et al., 2025b) proposes pattern-aware reasoning for deepfake detection, where the model walks through fast judgement, planning, reasoning, self-reflection, and conclusion stages before committing to a verdict. Trained on InternVL3-8B via pattern-guided cold-start and Pattern-Aware GRPO, it aims to deliver faithful decision traces with strong out-of-domain generalization.
- **Deep-VRM** (Lin et al., 2026) targets full-spectrum forensic signal perception in MLLMs, motivated by the observation that forcing early-to-middle LLM layers to learn low-level artifacts disrupts their semantic representations. It preserves early semantic processing and instead injects artifact-specific visual features as a residual path into an intermediate layer, where they are fused with semantic tokens and propagated through subsequent trainable layers without relying on any external detector.

Table 5: Implementation details of CounterFake.

Setting	Value
Model	
Base VLM	LLaVA-1.5-7B
Trainable parameters	LoRA on the vision encoder and LLM
LoRA rank / scaling factor	128 / 256
Training	
Training data	44,744
Optimizer	AdamW
Learning rate	5×10^{-6}, cosine schedule
Weight decay	0.01
Batch size	64
Epochs	4
Hardware	NVIDIA B200
Post-hoc robustness	
Weight interpolation alpha	0.5
Linear probe	Ridge on layer-11 hidden state, λ=10^3
Probe augmentation	Downscale (0.6–0.9×), JPEG (q 40–80), $p = 0.5$

- **TranX-Adapter** (Wang et al., 2026) is a lightweight fusion adapter that bridges texture-level artifact features and semantic features within MLLMs for AI-generated image detection. To counter the attention dilution caused by highly self-similar artifact features, it pairs a Task-aware Optimal-Transport Fusion, which uses the Jensen–Shannon divergence between artifact and semantic predictions as the transport cost to inject artifact cues into semantic features, with an X-Fusion that transfers semantic information back into artifact features via cross-attention.
- **X^2-DFD** (Chen et al., 2026) builds an explainable and extendable MLLM-based deepfake detector through three stages. It first assesses which forgery-related features the MLLM can intrinsically detect. It then constructs an explanation dataset by strengthening these strong features and supplementing weak ones with plug-and-play specific feature detectors. Finally, it fine-tunes the MLLM on this dataset for joint detection and explanation.
- **VERITAS++** (Tan et al., 2026) extends VERITAS from face forgery to general AI-generated image detection, attributing the remaining failures of MLLM detectors to perception bottlenecks rather than reasoning. Built on Qwen3-VL-8B, it combines a human-annotated cold start with two further stages. Perception-oriented Learning uses verifiable rewards to strengthen fine-grained, semantic-anomaly, and pixel-level perception. Value-aware On-Policy Distillation then internalizes perception-aware reasoning through a privileged self-teacher.

Generic MLLM. We also evaluate generic MLLMs, including LLaVA-1.5-7B (Liu et al., 2024), InternVL3-8B (Zhu et al., 2025), and GPT-4o (Hurst et al., 2024), Qwen2.5-VL Bai et al. (2025) which are queried in a zero-shot manner with carefully designed instruction prompts that cast deepfake detection as a visual question answering task, following the prompting protocol of Pirogov (2025).

C.3 Fine-tuning Details of CounterFake

We use LLaVA-1.5 with a 7B backbone for all experiments. LoRA is applied with a rank of 128 and a scaling factor $\alpha = 256$. For reproducibility, we provide the full implementation and training details of CounterFake in Table 5.

C.4 Post-hoc Model Selection and Inference Details

The ridge probe is fitted to binary labels on training representations standardized using feature-wise statistics from training samples. During probe fitting, we employ image augmentations consisting

of downscaling (0.6–0.9×) and JPEG compression (quality 40–80) with probability $p = 0.5$. The probing layer and ridge regularization strength are selected on the validation split described in Appendix C.1 and held fixed for test evaluation. At inference, the predicted label is inserted at the label position before rationale generation resumes. The layer-by-regularization validation sweep is shown in Figure 2b.

D Extended Benchmark Results

D.1 Per-method Cross-domain Results

Per-forgery results. Tables 6 and 7 report per-subset accuracy and macro-F1 in the cross-domain setting, where all 32 forgery subsets are evaluated with Celeb-DF as the unseen real-source domain. Since macro-F1 accounts for the F1-scores of both the real and fake classes, it better reveals single-class collapse, which can be obscured when considering accuracy alone. We therefore report macro-F1 alongside accuracy and use it as a key metric for evaluating cross-domain robustness. CounterFake achieves the best average performance on both metrics and consistently ranks among the strongest methods across the 32 forgery subsets. The close agreement between its accuracy and macro-F1 also suggests balanced predictions across the real and fake classes.

In contrast, several baselines show considerable variation in performance across forgery types, performing strongly on selected subsets but less consistently on others. This trend is particularly noticeable for some MLLM-based detectors, whose performance varies substantially across manipulation categories. LLaVA1.5-7B, Qwen2.5-VL-7B, and VERITAS also obtain relatively low macro-F1 scores on many subsets, suggesting less balanced predictions between the real and fake classes.

Among the competing methods, FakeVLM provides the strongest overall baseline and shares the same backbone as CounterFake. However, its performance degrades substantially from FS to FR, whereas CounterFake remains comparatively consistent across the two categories. This suggests that CounterFake is more robust to variations in manipulation type rather than relying on cues specific to particular forgery techniques. Nevertheless, audio-driven FR manipulations, such as sadtalker and wav2lip, remain challenging across all methods, including ours.

D.2 Per-Benchmark Results under Cross-domain & Forgery Shifts

Accuracy on cross-domain & forgery setting. Table 8 reports the accuracy under the cross-domain & forgery setting, complementing the macro-F1 results in Table 2. CounterFake achieves the highest average accuracy of 0.8843, outperforming the second-best method, FakeVLM, by 7.6 percentage points. Although several detectors perform strongly on individual benchmarks, their performance varies substantially across benchmarks. For example, Deep-VRM achieves the best performance on StyleCLIP but performs substantially worse on StarGAN and StarGANv2. Similarly, TranX-Adapter achieves the best performance on CollabDiff but shows a sharp performance drop on HeyGen. In contrast, CounterFake exceeds 0.81 accuracy on all seven benchmarks. Consistent with the macro-F1 results, these findings demonstrate that CounterFake provides stable detection performance even when both the real source domain and the forgery technique are unseen.

E Comparison with Vision-Only Detectors

We train five vision-only detectors on the same images used for CounterFake and select their checkpoints using the validation protocol in Appendix C.1.

Table 9 summarizes the results across the three evaluation settings. NPR achieves the highest in-domain accuracy at 0.979, compared with 0.909 for CounterFake. However, NPR suffers substantial performance degradation under distribution shifts, dropping by 47.7 percentage points on Cross-domain and 39.3 percentage points on Cross-domain & forgery. In contrast, CounterFake shows much smaller drops of 9.8 and 2.5 percentage points, respectively. This result highlights a trade-off between in-domain performance and robustness: while CounterFake performs slightly worse than NPR in-domain, it maintains substantially more stable performance under unseen domains and forgery techniques.

Table 6: **Per-subset accuracy on cross-domain benchmarks.** All 32 subsets use Celeb-DF (Li et al., 2020) as the unseen real source domain and are grouped by forgery type as in Table 1. The best and second-best results in each row are **bolded** and underlined, respectively.

	Generic MLLM				MLLM-based detector								
Subset	LLaVA1.5-7B	Qwen2.5-VL-7B	InternVL3-8B	GPT-4o	M2F2Det (7B)	Deep-VRM (8B)	TranX-Adapter (7B)	SIDA-13B	X^2-DFD (7B)	FakeVLM (7B)	VERITAS (8B)	VERITAS++ (8B)	**CounterFake (7B)**
FS (Face Swapping)													
blendface	0.5043	0.4994	0.5209	0.5431	0.5096	0.5082	0.5174	0.4994	0.6918	0.7419	0.4443	0.5043	**0.7595**
e4s	0.5278	0.5090	0.5809	0.5385	0.6575	0.5548	0.5051	0.7470	0.7110	0.7979	0.4388	0.6152	**0.8827**
facedancer	0.5037	0.5000	0.5189	0.5531	0.5417	0.4807	0.5019	0.6292	0.6663	0.7576	0.4110	0.5467	**0.8702**
faceswap	0.5422	0.5043	0.6883	0.5542	0.7912	0.5353	0.5006	0.5325	0.7218	0.8401	0.4797	0.7368	**0.8870**
fsgan	0.5079	0.5037	0.5507	0.5487	0.5572	0.5032	0.5000	0.5302	0.6895	0.8078	0.4552	0.5653	**0.8417**
inswap	0.5047	0.4991	0.5124	0.5369	0.5038	0.5192	0.5148	0.4994	0.6825	**0.8098**	0.4386	0.4920	0.5954
mobileswap	0.5035	0.4996	0.5000	0.5472	0.5153	0.4919	0.4990	0.5004	0.6829	0.7310	0.4446	0.5024	**0.8142**
simswap	0.5043	0.4996	0.5237	0.5488	0.5111	0.5225	0.5207	0.5033	0.6943	0.6681	0.4638	0.5078	**0.7097**
uniface	0.5042	0.4994	0.5260	0.5459	0.5176	0.5192	0.5138	0.5058	0.7027	0.7798	0.4457	0.5200	**0.7977**
FR (Face Reenactment)													
danet	0.5012	0.5004	0.5188	0.5368	0.5058	0.5769	0.4994	0.5141	0.5758	0.4461	0.4120	0.5977	**0.8568**
facevid2vid	0.5004	0.4994	0.5063	0.5560	0.5010	0.5919	0.4980	0.5073	0.6527	0.5744	0.3958	0.6126	**0.8830**
fomm	0.5052	0.5000	0.5659	0.5550	0.5131	0.5898	0.4971	0.5070	0.6188	0.6833	0.4318	0.6828	**0.8750**
hyperreenact	0.5002	0.5004	0.5037	0.5536	0.5016	0.5991	0.4994	0.5008	0.6544	0.6529	0.4239	0.5437	**0.8830**
lia	0.5023	0.5017	0.5442	0.5505	0.5037	0.5951	0.4990	0.5004	0.6533	**0.8023**	0.4263	0.5440	0.6725
mcnet	0.5002	0.5014	0.5163	0.5493	0.5039	0.5774	0.4996	0.5134	0.5927	0.4762	0.4005	0.6053	**0.8607**
MRAA	0.5025	0.4998	0.5253	0.5521	0.5127	0.5517	0.4979	0.4977	0.6800	**0.8234**	0.4325	0.5749	0.7673
one_shot_free	0.5091	0.5000	0.5632	0.5550	0.5141	0.5900	0.4977	0.5033	0.6065	0.6459	0.4301	0.6711	**0.8657**
pirender	0.5026	0.4994	0.5116	0.5521	0.5112	0.5680	0.4976	0.5024	0.5743	0.5963	0.4206	0.5888	**0.8420**
sadtalker	0.5032	0.5011	0.5159	0.5322	0.5040	0.4839	0.4977	0.4972	0.4714	0.4843	0.4295	0.5256	**0.5494**
tpsm	0.5004	0.5008	0.5118	0.5358	0.4981	0.4176	0.4971	0.5201	0.5362	0.5460	0.4120	0.5182	**0.8033**
wav2lip	0.5030	0.4994	0.5103	0.5150	0.5017	0.5087	0.5087	0.4987	0.5634	**0.5666**	0.4283	0.4880	0.4620
EFS (Entire Face Synthesis)													
ddim	0.5368	0.5210	0.5479	0.5514	0.5314	0.6633	0.8755	0.6639	0.7228	0.7211	0.3705	0.6208	**0.8859**
DiT	0.5049	0.4996	0.5234	0.5507	0.5228	0.7067	**0.8213**	0.5014	0.6890	0.6923	0.4240	0.5349	0.7175
pixart	0.5504	0.6032	0.7135	0.5514	0.6812	0.7105	**0.9951**	0.7565	0.7237	0.8355	0.4842	0.6225	0.8871
rddm	0.6006	0.5358	0.8873	0.5408	0.5513	0.3482	0.4963	0.5654	0.7217	0.8411	0.3935	**0.9051**	0.8856
sd2.1	0.5175	0.5242	0.6236	0.5514	0.6164	0.7123	**0.9680**	0.6029	0.7237	0.8313	0.4444	0.6076	0.8854
SiT	0.5062	0.4999	0.5313	0.5500	0.5244	0.7077	**0.8469**	0.5031	0.7015	0.7398	0.4342	0.5442	0.7619
StyleGAN2	0.5087	0.5000	0.5410	0.5514	0.6426	0.7125	0.5545	0.7332	0.7237	0.8355	0.4544	0.5917	**0.8886**
StyleGAN3	0.5248	0.5001	0.5494	0.5514	0.5679	0.7109	0.5742	0.5527	0.7236	0.8355	0.4495	0.5746	**0.8885**
StyleGANXL	0.5082	0.4996	0.5200	0.5514	0.6954	0.7074	0.5555	0.5362	0.7153	0.8353	0.4387	0.4970	**0.8840**
VQGAN	0.6195	0.5056	0.6907	0.5514	0.8044	0.7119	**0.9948**	0.8067	0.7236	0.8354	0.4725	0.7646	0.8885
FE (Facial Editing)													
e4e	0.5036	0.5014	0.4828	0.5507	0.6424	0.7075	0.5793	**0.9556**	0.7200	0.8366	0.4096	0.5465	0.8891
Avg.†	0.5161	0.5065	0.5571	0.5472	0.5611	0.5839	0.5883	0.5683	0.6660	0.7210	0.4325	0.5860	**0.8107**

†Avg. denotes the mean over all 32 forgery methods, identical to the Avg. column of Table 1.

E.1 Cross-domain Results

Tables 12 and 13 report performance across all 32 forgery methods under the Cross-domain setting. CounterFake achieves an average accuracy of 0.811 and a macro-F1 of 0.807. As summarized in Table 9, the best vision-only baseline achieves 0.625 accuracy with D^3 and 0.596 macro-F1 with UnivFD. CounterFake therefore improves over the strongest baseline for each metric by 18.6 percentage points in accuracy and 21.1 percentage points in macro-F1.

This result contrasts with the in-domain setting in Table 9, where several vision-only baselines outperform CounterFake. Under the Cross-domain setting, however, these baselines exhibit substantially larger performance degradation, whereas CounterFake maintains considerably higher accuracy and macro-F1 across the 32 forgery methods. These results indicate that CounterFake is substantially more robust to shifts in the real-source domain.

E.2 Cross-domain & Forgery Results

Tables 10 and 11 report accuracy and macro-F1 across the seven forgery methods in the Cross-domain & forgery setting. Among the vision-only baselines, UnivFD achieves the strongest average performance, with an accuracy of 0.692 and a macro-F1 of 0.646. In comparison, CounterFake

Table 7: **Per-subset macro-F1 on cross-domain benchmarks.** All 32 subsets use Celeb-DF (Li et al., 2020) as the unseen real source domain and are grouped by forgery type as in Table 1. The best and second-best results in each row are **bolded** and underlined, respectively.

	Generic MLLM				MLLM-based detector								
Subset	LLaVA1.5-7B	Qwen2.5-VL-7B	InternVL3-8B	GPT-4o	M2F2Det (7B)	Deep-VRM (8B)	TranX-Adapter (7B)	SIDA-13B	X^2-DFD (7B)	FakeVLM (7B)	VERITAS (8B)	VERITAS++ (8B)	**CounterFake (7B)**
FS (Face Swapping)													
blendface	0.3428	0.3331	0.4297	0.4325	0.3598	0.5050	0.3752	0.3422	0.6718	0.7411	0.3464	0.4050	**0.7594**
e4s	0.3925	0.3551	0.5279	0.4135	0.6142	0.5452	0.3497	0.7315	0.6880	0.7962	0.3557	0.5805	**0.8811**
facedancer	0.3414	0.3344	0.4263	0.4415	0.4244	0.4789	0.3426	0.5743	0.6488	0.7564	0.3295	0.4761	**0.8691**
faceswap	0.4208	0.3439	0.6704	0.4449	0.7825	0.5302	0.3395	0.4098	0.6988	0.8362	0.3679	0.7284	**0.8856**
fsgan	0.3508	0.3425	0.4796	0.4433	0.4532	0.5003	0.3380	0.4055	0.6698	0.8050	0.3524	0.5060	**0.8410**
inswap	0.3437	0.3329	0.4178	0.4203	0.3470	0.5141	0.3714	0.3428	0.6628	**0.8073**	0.3441	0.3844	0.5822
mobileswap	0.3412	0.3335	0.3923	0.4414	0.3718	0.4900	0.3360	0.3441	0.6638	0.7305	0.3470	0.4016	**0.8139**
simswap	0.3428	0.3335	0.4343	0.4408	0.3631	0.5179	0.3820	0.3506	0.6740	0.6680	0.3611	0.4112	**0.7084**
uniface	0.3426	0.3331	0.4386	0.4331	0.3762	0.5146	0.3677	0.3557	0.6815	0.7778	0.3509	0.4325	**0.7976**
FR (Face Reenactment)													
danet	0.3359	0.3352	0.4258	0.4304	0.3517	0.5671	0.3372	0.3732	0.5683	0.4128	0.3334	0.5531	**0.8559**
facevid2vid	0.3342	0.3331	0.4037	0.4505	0.3413	0.5801	0.3338	0.3587	0.6370	0.5689	0.3258	0.5742	**0.8817**
fomm	0.3449	0.3344	0.5033	0.4463	0.3676	0.5791	0.3320	0.3582	0.6067	0.6833	0.3424	0.6649	**0.8738**
hyperreenact	0.3338	0.3352	0.3989	0.4425	0.3426	0.5867	0.3368	0.3451	0.6385	0.6525	0.3426	0.4716	**0.8817**
lia	0.3387	0.3382	0.4686	0.4403	0.3471	0.5832	0.3363	0.3442	0.6374	**0.7995**	0.3381	0.4720	0.6688
mcnet	0.3338	0.3374	0.4215	0.4452	0.3475	0.5670	0.3376	0.3715	0.5834	0.4524	0.3231	0.5638	**0.8597**
MRAA	0.3389	0.3339	0.4372	0.4397	0.3663	0.5443	0.3337	0.3383	0.6611	**0.8200**	0.3385	0.5197	0.7672
one_shot_free	0.3536	0.3344	0.4991	0.4463	0.3692	0.5786	0.3333	0.3504	0.5957	0.6453	0.3424	0.6505	**0.8647**
pirender	0.3390	0.3331	0.4133	0.4409	0.3633	0.5591	0.3330	0.3483	0.5670	0.5929	0.3316	0.5405	**0.8413**
sadtalker	0.3404	0.3369	0.4208	0.4274	0.3480	0.4823	0.3333	0.3377	0.4710	0.4632	0.3379	0.4420	**0.5251**
tpsm	0.3342	0.3361	0.4136	0.4297	0.3349	0.4176	0.3320	0.3853	0.5321	0.5371	0.3316	0.4288	**0.8032**
wav2lip	0.3400	0.3331	0.4108	0.4182	0.3430	0.5054	0.3570	0.3409	0.5569	**0.5607**	0.3401	0.3752	0.4033
EFS (Entire Face Synthesis)													
ddim	0.4102	0.3790	0.4747	0.4386	0.4040	0.6424	0.8737	0.6242	0.7000	0.7204	0.3075	0.5854	**0.8846**
DiT	0.3444	0.3331	0.4338	0.4381	0.3871	0.6810	**0.8158**	0.3461	0.6697	0.6922	0.3392	0.4574	0.7165
pixart	0.4365	0.5294	0.7050	0.4386	0.6483	0.6842	**0.9951**	0.7427	0.7009	0.8310	0.3661	0.5877	0.8857
rddm	0.5306	0.4094	0.8872	0.4182	0.4422	0.3445	0.3317	0.4710	0.6983	0.8374	0.3259	**0.9051**	0.8841
sd2.1	0.3714	0.3856	0.5871	0.4386	0.5537	0.6867	**0.9680**	0.5335	0.7009	0.8269	0.3435	0.5672	0.8840
SiT	0.3473	0.3338	0.4472	0.4377	0.3901	0.6823	**0.8435**	0.3497	0.6809	0.7386	0.3462	0.4725	0.7618
StyleGAN2	0.3525	0.3341	0.4635	0.4386	0.5926	0.6866	0.4474	0.7145	0.7009	0.8310	0.3527	0.5447	**0.8872**
StyleGAN3	0.3874	0.3344	0.4772	0.4386	0.4723	0.6846	0.4828	0.4476	0.7008	0.8310	0.3509	0.5195	**0.8871**
StyleGANXL	0.3513	0.3331	0.4279	0.4386	0.6659	0.6819	0.4493	0.4167	0.6933	0.8307	0.3448	0.3918	**0.8826**
VQGAN	0.5553	0.3465	0.6733	0.4386	0.7986	0.6859	**0.9948**	0.8000	0.7008	0.8308	0.3617	0.7593	0.8871
FE (Facial Editing)													
e4e	0.3413	0.3373	0.3590	0.4371	0.5923	0.6826	0.4917	**0.9555**	0.6975	0.8322	0.3257	0.4756	0.8878
Avg.†	0.3661	0.3475	0.4803	0.4366	0.4457	0.5715	0.4666	0.4534	0.6487	0.7159	0.3421	0.5265	**0.8067**

†Avg. denotes the mean over all 32 forgery methods, identical to the Avg. column of Table 1.

achieves 0.884 accuracy and 0.881 macro-F1, outperforming the strongest baseline by 19.2 and 23.5 percentage points, respectively.

The improvement is consistent across forgery methods. As shown in Table 10, CounterFake achieves the highest accuracy on all seven methods. Table 11 shows a similarly consistent trend for macro-F1: CounterFake achieves the highest score on six of the seven methods and ties with UnivFD on HeyGen. The gains are particularly pronounced on CollabDiff, StarGAN, StarGANv2, and StyleCLIP, where the vision-only baselines show substantial performance degradation under the combined domain and forgery shift. These results demonstrate that CounterFake generalizes more consistently when both the real-source domain and forgery method are unseen during training.

F Additional Ablations, Generalization, and Robustness Analysis

F.1 Inference efficiency

We analyze the inference latency of CounterFake compared to Effort (Yan et al., 2024a), a CLIP-based vision-only baseline, as presented in Table 14. While producing full rationales increases total latency to 0.0130s, CounterFake remains practically efficient through modern inference optimizations. By leveraging vLLM with KV-caching and batching, our model achieves a high throughput of

Table 8: **Accuracy on cross-domain & forgery benchmarks.** All evaluation datasets are generated from unseen real source domains using unseen forgery techniques, neither of which were used during training. The best and second-best results are **bolded** and underlined, respectively.

Method	DeepFaceLab	HeyGen	WhichIsReal	CollabDiff	StarGAN	StarGANv2	StyleCLIP	Avg.
Generic MLLM								
LLaVA1.5-7B (Liu et al., 2024)	0.4402	0.4935	0.4948	0.4980	0.5955	0.5015	0.3743	0.4854
Qwen2.5-VL-7B (Bai et al., 2025)	0.4416	0.5081	0.4978	0.4970	0.5439	0.5010	0.3630	0.4789
InternVL3-8B (Zhu et al., 2025)	0.5627	0.7277	0.5252	0.4935	0.6512	0.5010	0.3820	0.5490
GPT-4o (Hurst et al., 2024)	0.5737	0.6932	0.7736	0.7390	0.5512	0.6081	0.8267	0.6808
MLLM-based detector								
M2F2Det (7B) (Guo et al., 2025)	0.5129	0.5065	0.5022	0.5065	0.7722	0.5210	0.4377	0.5370
Deep-VRM (8B) (Lin et al., 2026)	0.5240	0.7518	0.7581	0.6940	0.5199	0.4760	**0.9827**	0.6724
TranX-Adapter (7B) (Wang et al., 2026)	0.4160	0.3689	0.6702	**0.9810**	0.5847	0.4900	0.4073	0.5597
SIDA-13B (Huang et al., 2025)	0.4692	0.6409	0.5072	0.7580	0.7646	0.5971	0.9130	0.6643
X^2-DFD (7B) (Chen et al., 2026)	0.7289	0.6396	0.5097	0.5130	0.5522	0.5328	0.6643	0.5915
FakeVLM (7B) (Wen et al., 2025)	0.7962	0.6764	0.7966	0.8855	0.8863	0.6524	0.9627	0.8080
VERITAS (8B) (Tan et al., 2025b)	0.5260	0.7242	**0.9015**	0.8060	0.5628	0.4595	0.5367	0.6452
VERITAS++ (8B) (Tan et al., 2026)	0.5467	0.7531	0.6492	0.6820	0.7588	0.5280	0.7283	0.6637
CounterFake (7B)	**0.8829**	**0.8108**	0.9010	0.9035	**0.9110**	**0.8493**	0.9317	**0.8843**

Table 9: **Overall results of vision-only detectors.** We report accuracy (Acc.) and macro-F1 (mF1) on the in-domain, cross-domain, and cross-domain & forgery benchmarks. The best and second-best results in each column are **bolded** and underlined, respectively.

	In-domain		Cross-domain		Cross-domain & forgery	
Method	Acc.	mF1	Acc.	mF1	Acc.	mF1
UnivFD	0.892	0.891	0.624	0.596	0.692	0.646
NPR	**0.979**	**0.979**	0.502	0.343	0.586	0.457
D^3	0.919	0.918	0.625	0.586	0.667	0.591
Co-SPY	0.945	0.945	0.467	0.330	0.540	0.401
Effort	0.970	0.970	0.521	0.395	0.677	0.599
CounterFake	0.909	0.909	**0.811**	**0.807**	**0.884**	**0.881**

4,779.1 tokens per second for full generations. Since CounterFake produces the final verdict before the explanation, it can also operate in an "answer-only" mode when explanations are unnecessary, reducing generation overhead while retaining the option to provide detailed forensic rationales.

F.2 Comparison with Alternative Preference Objectives

Table 15 compares our preference loss with DPO (Rafailov et al., 2023), CHiP (Fu et al., 2025b), and MCM-DPO (Fu et al., 2025a) under the same SFT baseline and training setup. From an SFT average accuracy of 0.8479, DPO remains nearly unchanged at 0.8466, while CHiP and MCM-DPO improve to 0.8627 and 0.8766. Our loss achieves the highest average accuracy of 0.8843, exceeding MCM-DPO by 0.77 percentage points, and performs best on five of seven benchmarks.

F.3 Effect of Hard-negative Candidate Pool Size

As shown in Table 16, we investigate the effect of the top-k sampling scheme on the efficiency of hard-negative retrieval. As k increases, Hit@1 and Hit@10 consistently improve, indicating that a larger candidate pool increases the likelihood of selecting visually similar and semantically confusing negatives. Notably, the performance gain from $k = 100$ to $k = 500$ is relatively modest compared to the significant increase in computation time (13.24 ms *vs.* 65.95 ms). Based on this diminishing return, we set $k = 100$ in all experiments as a practical trade-off between accuracy and efficiency.

Table 10: **Accuracy of vision-only detectors on cross-domain & forgery benchmarks.** We evaluate the detectors on the seven test sets in Table 8. Avg. is the unweighted mean across sets. The best and second-best results in each column are **bolded** and underlined, respectively.

Method	DeepFaceLab	HeyGen	WhichIsReal	CollabDiff	StarGAN	StarGANv2	StyleCLIP	Avg.
Vision-only detectors								
UnivFD (Ojha et al., 2023)	0.803	0.809	0.689	0.604	0.622	0.603	0.715	0.692
NPR (Tan et al., 2024b)	0.578	0.691	0.504	0.576	0.511	0.511	0.729	0.586
D^3 (Yang et al., 2025)	0.821	0.767	0.708	0.562	0.560	0.549	0.702	0.667
Co-SPY (Cheng et al., 2025)	0.519	0.535	0.504	0.582	0.449	0.468	0.725	0.540
Effort (Yan et al., 2024a)	0.804	0.737	0.811	0.607	0.516	0.516	0.746	0.677
CounterFake	**0.883**	**0.811**	**0.901**	**0.903**	**0.911**	**0.849**	**0.932**	**0.884**

Table 11: **Macro-F1 of vision-only detectors on cross-domain & forgery benchmarks.** We use the same seven test sets as in Table 10. The best and second-best results in each column are **bolded** and underlined, respectively.

Method	DeepFaceLab	HeyGen	WhichIsReal	CollabDiff	StarGAN	StarGANv2	StyleCLIP	Avg.
Vision-only detectors								
UnivFD (Ojha et al., 2023)	0.800	**0.806**	0.685	0.541	0.560	0.535	0.593	0.646
NPR (Tan et al., 2024b)	0.402	0.654	0.342	0.484	0.380	0.366	0.573	0.457
D^3 (Yang et al., 2025)	0.816	0.764	0.691	0.461	0.455	0.433	0.518	0.591
Co-SPY (Cheng et al., 2025)	0.355	0.393	0.346	0.498	0.322	0.326	0.566	0.401
Effort (Yan et al., 2024a)	0.782	0.720	0.806	0.534	0.367	0.367	0.612	0.599
CounterFake	**0.882**	**0.806**	**0.901**	**0.903**	**0.910**	**0.849**	**0.919**	**0.881**

F.4 Hard-negative Retrieval Design

Since LLaVA-1.5 uses a CLIP-based image encoder, we first retrieve candidate rejected images using CLIP and then select hard negatives based on likelihood. To examine the effect of this retrieval strategy, we compare it with CLIP, DINOv2, and ImageReward, training all variants on the same 44K training set while keeping the inference pipeline fixed. As shown in Table 17, our CLIP + likelihood retrieval consistently performs best across both evaluation settings, achieving the highest average accuracy. These results suggest that likelihood-based hard-negative selection provides a more effective way to identify informative rejected images than relying on embedding-based retrieval alone.

F.5 Training Stability of Hard-negative Mining

We further compare the training stability of our likelihood-based hard-negative retrieval with three alternatives: (i) DINOv2-based retrieval, (ii) CLIP-only, and (iii) ImageReward-based retrieval. As shown in Figure 10, our method exhibits slightly larger variations in gradient norm than the alternatives, while remaining stable throughout training without signs of divergence. Together with the performance gains over CLIP-only retrieval reported in Table 17, these results suggest that likelihood-based hard-negative selection improves retrieval effectiveness without compromising training stability.

F.6 Feature-space Analysis

To further investigate the source of CounterFake's strong generalization, we visualize the feature space using t-SNE under the cross-domain & forgery setting. As shown in Figure 11, while pretrained CLIP yields heavily entangled real and fake embeddings across samples from all seven unseen domains and forgeries, CounterFake produces more distinctly separated clusters, indicating that our preference optimization induces a more discriminative manifold.

F.7 Backbone Generalization

Results on Qwen-2.5-VL-3B backbone. To verify that the effectiveness of CounterFake is not restricted to a specific VLM, *i.e.*, LLaVA-1.5, we additionally test CounterFake on Qwen-2.5-VL-

Table 12: **Per-subset accuracy of vision-only detectors on cross-domain benchmarks.** All 32 subsets use Celeb-DF (Li et al., 2020) as the unseen real source domain and follow the grouping in Table 6. The best and second-best results in each row are **bolded** and underlined, respectively.

Subset	UnivFD	NPR	D^3	Co-SPY	Effort	**CounterFake**
FS						
blendface	0.588	0.503	0.620	0.486	0.529	**0.7595**
e4s	0.666	0.503	0.645	0.502	0.527	**0.8827**
facedancer	0.659	0.502	0.647	0.475	0.532	**0.8702**
faceswap	0.674	0.505	0.657	0.452	0.532	**0.8870**
fsgan	0.648	0.504	0.640	0.457	0.531	**0.8417**
inswap	0.551	0.494	0.555	0.381	0.481	**0.5954**
mobileswap	0.631	0.505	0.630	0.496	0.532	**0.8142**
simswap	0.587	0.502	0.613	0.483	0.523	**0.7097**
uniface	0.607	0.505	0.615	0.505	0.531	**0.7977**
FR						
danet	0.607	0.505	0.639	0.487	0.532	**0.8568**
facevid2vid	0.617	0.505	0.648	0.371	0.532	**0.8830**
fomm	0.658	0.505	0.651	0.481	0.532	**0.8750**
hyperreenact	0.674	0.503	0.657	0.501	0.532	**0.8830**
lia	0.535	0.485	0.542	0.405	0.438	**0.6725**
mcnet	0.610	0.504	0.637	0.489	0.532	**0.8607**
MRAA	0.594	0.497	0.606	0.336	0.505	**0.7673**
one_shot_free	0.652	0.505	0.641	0.486	0.532	**0.8657**
pirender	0.645	0.505	0.639	0.492	0.532	**0.8420**
sadtalker	0.560	0.504	**0.575**	0.502	0.529	0.5494
tpsm	0.566	0.505	0.627	0.458	0.532	**0.8033**
wav2lip	0.499	**0.500**	0.499	0.496	0.438	0.4620
EFS						
ddim	0.681	0.500	0.661	0.485	0.532	**0.8859**
DiT	0.510	0.503	0.550	0.499	0.498	**0.7175**
pixart	0.682	0.505	0.661	0.485	0.532	**0.8871**
rddm	0.674	0.505	0.651	0.508	0.531	**0.8856**
sd2.1	0.674	0.503	0.658	0.489	0.532	**0.8854**
SiT	0.510	0.504	0.553	0.500	0.507	**0.7619**
StyleGAN2	0.684	0.505	0.659	0.481	0.532	**0.8886**
StyleGAN3	0.682	0.505	0.659	0.469	0.532	**0.8885**
StyleGANXL	0.681	0.493	0.660	0.283	0.532	**0.8840**
VQGAN	0.682	0.504	0.659	0.505	0.532	**0.8885**
FE						
e4e	0.681	0.504	0.660	0.509	0.533	**0.8891**
Avg.	0.624	0.502	0.625	0.467	0.521	**0.8107**

3B, which adopts a different architecture and vision encoder. Being smaller in scale (3B *vs.* 7B), Qwen-2.5-VL-3B provides a distinct backbone for evaluating the generalizability of our approach. As shown in Table 18, CounterFake improves the average accuracy over SFT from 0.6342 to 0.7086 under the challenging cross-domain & forgery setting. This confirms that our approach is not tied to a specific VLM backbone.

F.8 Robustness to Common Image Perturbations

Robustness analysis. We further evaluate model robustness against five common image perturbations: Gaussian blur, Gaussian noise, contrast reduction, sharpening, and WebP compression, each under varying severity levels. As shown in Table 19, CounterFake consistently achieves the highest mF1 across all perturbation settings and obtains the best average mF1 of 0.8258, compared with 0.7169 for the strongest baseline, FakeVLM. Figure 12 further visualizes this robustness across

Table 13: **Per-subset macro-F1 of vision-only detectors on cross-domain benchmarks.** All 32 subsets use Celeb-DF (Li et al., 2020) as the unseen real source domain and follow the grouping in Table 7. The best and second-best results in each row are **bolded** and underlined, respectively.

Subset	UnivFD	NPR	D^3	Co-SPY	Effort	**CounterFake**
FS						
blendface	0.566	0.343	0.581	0.340	0.399	**0.7594**
e4s	0.625	0.341	0.593	0.344	0.391	**0.8811**
facedancer	0.626	0.343	0.603	0.335	0.401	**0.8691**
faceswap	0.638	0.344	0.612	0.323	0.401	**0.8856**
fsgan	0.617	0.344	0.596	0.325	0.400	**0.8410**
inswap	0.532	0.339	0.526	0.285	0.373	**0.5822**
mobileswap	0.602	0.344	0.588	0.345	0.401	**0.8139**
simswap	0.565	0.343	0.576	0.339	0.396	**0.7084**
uniface	0.581	0.344	0.577	0.349	0.400	**0.7976**
FR						
danet	0.582	0.344	0.598	0.340	0.400	**0.8559**
facevid2vid	0.590	0.344	0.603	0.280	0.401	**0.8817**
fomm	0.625	0.344	0.607	0.337	0.401	**0.8738**
hyperreenact	0.638	0.343	0.612	0.347	0.401	**0.8817**
lia	0.521	0.335	0.517	0.298	0.347	**0.6688**
mcnet	0.585	0.344	0.595	0.341	0.400	**0.8597**
MRAA	0.571	0.341	0.569	0.260	0.386	**0.7672**
one_shot_free	0.620	0.344	0.598	0.340	0.401	**0.8647**
pirender	0.613	0.344	0.597	0.343	0.401	**0.8413**
sadtalker	0.542	0.344	**0.545**	0.347	0.399	0.5251
tpsm	0.547	0.344	0.589	0.326	0.400	**0.8032**
wav2lip	**0.490**	0.342	0.482	0.344	0.346	0.4033
EFS						
ddim	0.646	0.341	0.616	0.340	0.402	**0.8846**
DiT	0.500	0.343	0.525	0.347	0.382	**0.7165**
pixart	0.647	0.343	0.617	0.341	0.402	**0.8857**
rddm	0.636	0.345	0.603	0.350	0.399	**0.8841**
sd2.1	0.640	0.343	0.614	0.343	0.401	**0.8840**
SiT	0.500	0.343	0.526	0.348	0.387	**0.7618**
StyleGAN2	0.649	0.343	0.614	0.339	0.402	**0.8872**
StyleGAN3	0.647	0.343	0.614	0.333	0.402	**0.8871**
StyleGANXL	0.647	0.338	0.615	0.228	0.402	**0.8826**
VQGAN	0.647	0.343	0.615	0.350	0.402	**0.8871**
FE						
e4e	0.646	0.343	0.615	0.352	0.403	**0.8878**
Avg.	0.596	0.343	0.586	0.330	0.395	**0.8067**

Table 14: **Per-sample inference latency.** Measured on 100 images over 10 runs using an NVIDIA B200 GPU.

Method	Inference Time (s)
Effort (Vision-only)	0.0025 ± 0.0001
CounterFake (Total)	0.0130 ± 0.0056
↪ Answer-only	0.0037 ± 0.0051
↪ Rationale	0.0093 ± 0.0076

varying perturbation severities, showing that CounterFake consistently maintains higher mF1 than the baselines. These results demonstrate that CounterFake maintains robust detection performance under diverse image degradations.

Table 15: **Accuracy comparison with multimodal preference optimization schemes.** We replace our preference learning objective with existing multimodal preference optimization methods, including DPO (Rafailov et al., 2023), CHiP (Fu et al., 2025b), and MCM-DPO (Fu et al., 2025a), and report accuracy on the cross-domain & forgery benchmarks.

Method	StarGAN	StarGANv2	StyleCLIP	CollabDiff	WhichIsReal	DeepFaceLab	HeyGen	**Avg.**
SFT	0.9062	0.8283	0.9123	0.8680	0.8226	0.8031	0.7946	0.8479
+ DPO	0.8843	0.8203	0.8623	0.8080	0.8596	0.8295	**0.8618**	0.8466
+ CHiP	0.8962	0.8283	0.9197	0.8875	0.8971	0.8684	0.7417	0.8627
+ MCM-DPO	0.8889	0.8343	0.9173	0.8910	0.8940	**0.8860**	0.8250	0.8766
+ Preference loss (Ours)	**0.9110**	**0.8493**	**0.9317**	**0.9035**	**0.9010**	0.8829	0.8108	**0.8843**

Table 16: **Ablation study on the effect of hard-negative pool size**. We randomly sample 1,000 real and fake training examples, and conduct all experiments on a single NVIDIA A40 GPU.

Top-k	Hit@10 (%)	Hit@1 (%)	Time (ms)
10	45.3	1.2	1.41 ± 0.12
50	81.9	9.2	6.62 ± 0.22
100	90.8	17.9	13.24 ± 0.44
500	98.7	31.5	65.95 ± 2.09

G Forensic Rationale and Caption Quality Evaluation

The main-paper reliability analysis compares ReForensics with single-image standard captions on 500 randomly sampled DF40 images, using multiple evaluation MLLMs. Figure 16 gives the scoring prompt and its four stated criteria; the resulting scores are shown in Figure 5 of the main paper.

We further provide the details of the two human studies conducted on 20 fake image samples each. The first study (Survey 1), conducted with 32 participants, evaluates the forensic rationales produced by CounterFake, and its results are reported in Figure 3 of the main paper. The second study (Survey 2), conducted with 31 participants, assesses the caption quality of ReForensics. To mitigate ordering bias, we randomized both the sample order and the presentation order of the anonymized responses in each survey.

G.1 Survey 1: Model-generated Rationale Quality

Survey 1: Rationale quality comparison. To construct the evaluation set, we first randomly sampled 1,000 fake images from DF40 (Yan et al., 2024b) and identified the samples correctly classified by all four models: CounterFake, VERITAS (Tan et al., 2025b), VERITAS++ (Tan et al., 2026), and FakeVLM (Wen et al., 2025). From this jointly correct subset, we randomly selected 20 samples and collected the rationales generated by each model. This setup allows the human evaluation to focus on rationale quality rather than classification correctness.

For each sample, participants ranked the four anonymized rationales from rank-1 (best) to rank-4 (worst) based on the following criteria:

- **Artifact localization:** correct identification of manipulated regions (e.g., eyes, jawline, hair, background).
- **Faithfulness:** absence of unsupported or incorrect claims.
- **Descriptive depth:** meaningful and non-redundant reasoning.

The user interface is shown in Figure 13. As reported in Figure 3, CounterFake receives the highest proportion of rank-1 ratings across all three criteria. Note that these results are limited to the 20 fake images correctly classified by all four models. Moreover, the ratings reflect the quality of the rationales as perceived by participants, rather than whether each rationale causally reflects the model's decision.

Table 17: **Comparison of alternative image retrieval strategies.** Only the rejected image of each preference tuple is changed, and every model is read out with the same inference pipeline.

Method	Cross-domain		Cross-domain & forgery	
	Acc	mF1	Acc	mF1
CLIP	0.7997	0.7952	0.8561	0.8539
DINOv2	0.8079	0.8033	0.8751	0.8715
ImageReward	0.8036	0.7985	0.8719	0.8677
CLIP + likelihood retrieval (Ours)	**0.8107**	**0.8067**	**0.8843**	**0.8813**

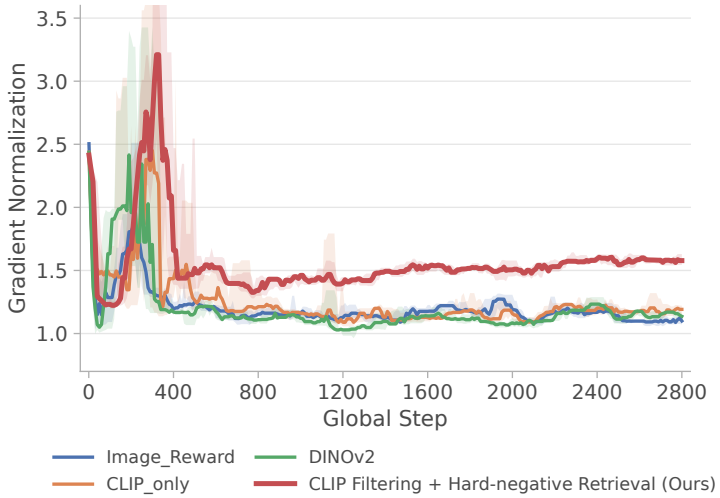


Figure 10: **Training stability of hard-negative retrieval schemes.** Gradient norm trajectories during CounterFake training under four retrieval schemes: (i) DINOv2, (ii) CLIP-only, (iii) ImageReward-based, and (iv) our likelihood-based hard-negative selection. While our scheme exhibits marginally higher variance reflecting the harder negatives it selects, training remains stable and consistently converges.

G.2 Survey 2: ReForensics Caption Quality

Survey 2: Caption quality comparison. To complement the machine-based evaluation in Figure 5, we conducted a human evaluation comparing ReForensics captions with standard captions. The standard captions follow the single-image prompt design of FakeVLM (Wen et al., 2025), whereas ReForensics additionally incorporates reference-guided comparison and forensic expertise prompts derived from our artifact taxonomy. We randomly sampled 20 fake images from DF40 (Yan et al., 2024b) and collected both captions for each image, allowing participants to directly compare caption quality under the same visual input.

For each sample, participants chose the better of the two anonymized captions separately under each of the following criteria:

- **Artifact localization:** accurate identification of manipulated regions.
- **Artifact explanation:** technically valid explanation of why the artifacts appear unnatural.
- **Description clarity:** specific and unambiguous language without vague claims.
- **Overall quality:** holistic reliability and forensic usefulness.

The user interface is shown in Figure 14. As reported in Figure 15, ReForensics captions are preferred over standard captions in more than 71% of cases across all four criteria. This indicates that, from a human perspective, ReForensics produces more reliable and well-grounded forensic captions, consistent with the machine-based reliability evaluation in Figure 5. Together with Figure 4,



Figure 11: **Impact of CounterFake on feature representation under the cross-domain & forgery setting. (Left)** With pre-trained CLIP, real and fake embeddings are heavily entangled across all seven unseen forgery types. **(Right)** With CounterFake, the feature space exhibits enhanced separability, forming a more discriminative manifold with distinct real and fake clusters.

Table 18: **Results on Qwen-2.5-VL-3B backbone.** We replace the LLaVA-1.5-7B backbone with Qwen-2.5-VL-3B, which adopts a different vision-language encoder, and report accuracy under the cross-domain & forgery setting. CounterFake improves the average accuracy over SFT alone.

Method	StarGAN	StarGANv2	StyleCLIP	CollabDiff	WhichIsReal	DeepFaceLab	HeyGen	Avg.
SFT	0.7258	**0.5901**	0.7887	0.6770	0.4888	**0.6897**	0.4796	0.6342
CounterFake	**0.7276**	0.5373	**0.8237**	**0.7995**	**0.6042**	0.6618	**0.8063**	**0.7086**

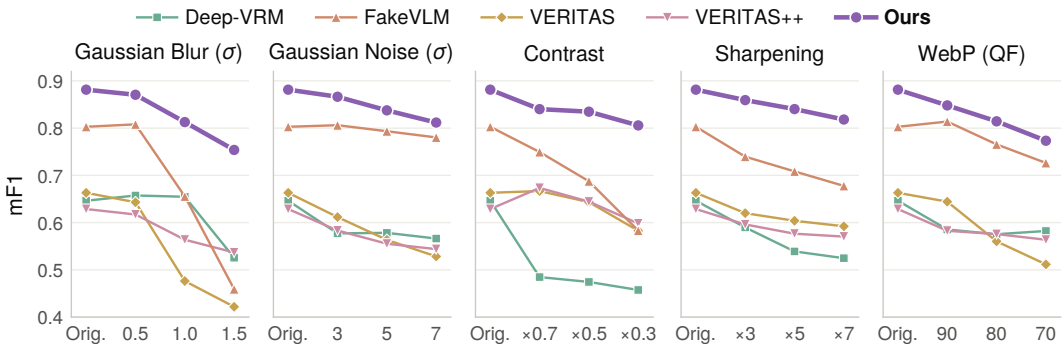


Figure 12: **Robustness to common image perturbations.** Set-average macro-F1 under the Cross-domain & forgery setting as perturbation strength varies across Gaussian blur, Gaussian noise, contrast reduction, sharpening, and WebP compression. For contrast and sharpening, $\times 1$ denotes the unperturbed image.

this suggests that ReForensics provides effective reasoning supervision for preference optimization in CounterFake.

The reliability scores and the participant preferences provide complementary assessments of ReForensics caption quality.

Table 19: **Robustness evaluation in mF1-Score against common perturbations.** We report the set-average macro-F1 under the Cross-domain & forgery setting across Gaussian blur, Gaussian noise, contrast reduction, sharpening, and WebP compression; contrast and sharpening are parameterized by enhancement factors ($\times 1$ leaves the image unchanged). The best results are **bolded**.

Method	Original	Gaussian Blur			Gaussian Noise			Contrast			Sharpening			WebP			Avg.
		σ=0.5	σ=1.0	σ=1.5	σ=3	σ=5	σ=7	$\times$0.7	$\times$0.5	$\times$0.3	$\times$3	$\times$5	$\times$7	QF=90	QF=80	QF=70	
Deep-VRM	0.6463	0.6574	0.6547	0.5257	0.5772	0.5783	0.5663	0.4847	0.4744	0.4574	0.5898	0.5390	0.5247	0.5854	0.5754	0.5823	0.5582
FakeVLM	0.8028	0.8080	0.6555	0.4591	0.8062	0.7934	0.7803	0.7494	0.6875	0.5831	0.7396	0.7084	0.6777	0.8139	0.7656	0.7263	0.7169
VERITAS	0.6631	0.6431	0.4763	0.4218	0.6117	0.5635	0.5284	0.6668	0.6434	0.5828	0.6200	0.6039	0.5922	0.6443	0.5603	0.5116	0.5780
VERITAS++	0.6291	0.6173	0.5638	0.5368	0.5833	0.5553	0.5439	0.6734	0.6444	0.5982	0.5961	0.5765	0.5704	0.5827	0.5757	0.5637	0.5854
CounterFake	**0.8813**	**0.8706**	**0.8128**	**0.7538**	**0.8664**	**0.8376**	**0.8118**	**0.8402**	**0.8348**	**0.8057**	**0.8593**	**0.8403**	**0.8182**	**0.8481**	**0.8144**	**0.7733**	**0.8258**

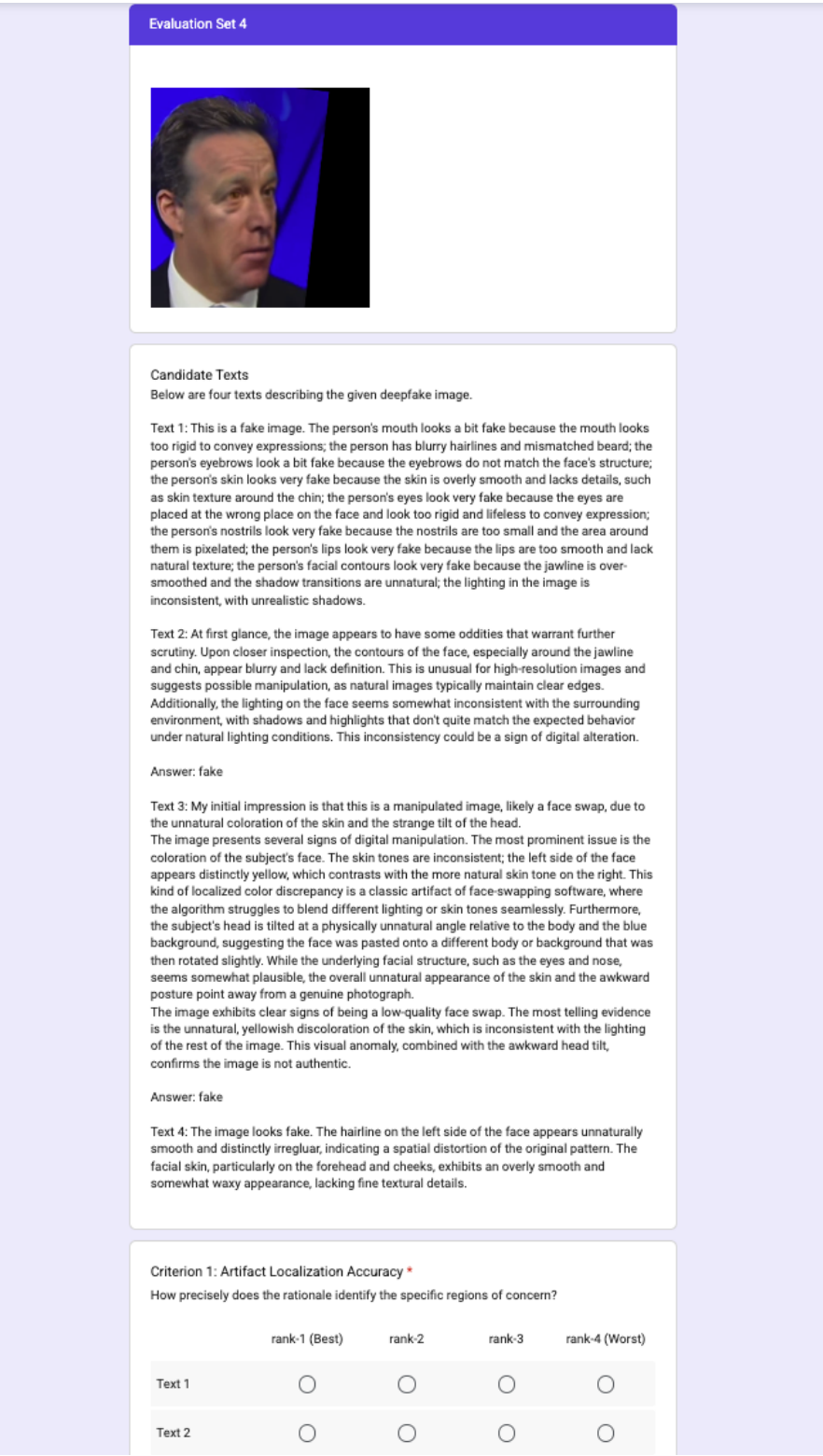


Figure 13: **User interface of the human evaluation (Survey 1).** For each forged image, participants rank four anonymized rationales (CounterFake, VERITAS, VERITAS++, and FakeVLM, presented in random order) from rank-1 (best) to rank-4 (worst) under three criteria. Each participant completes 20 such sets.

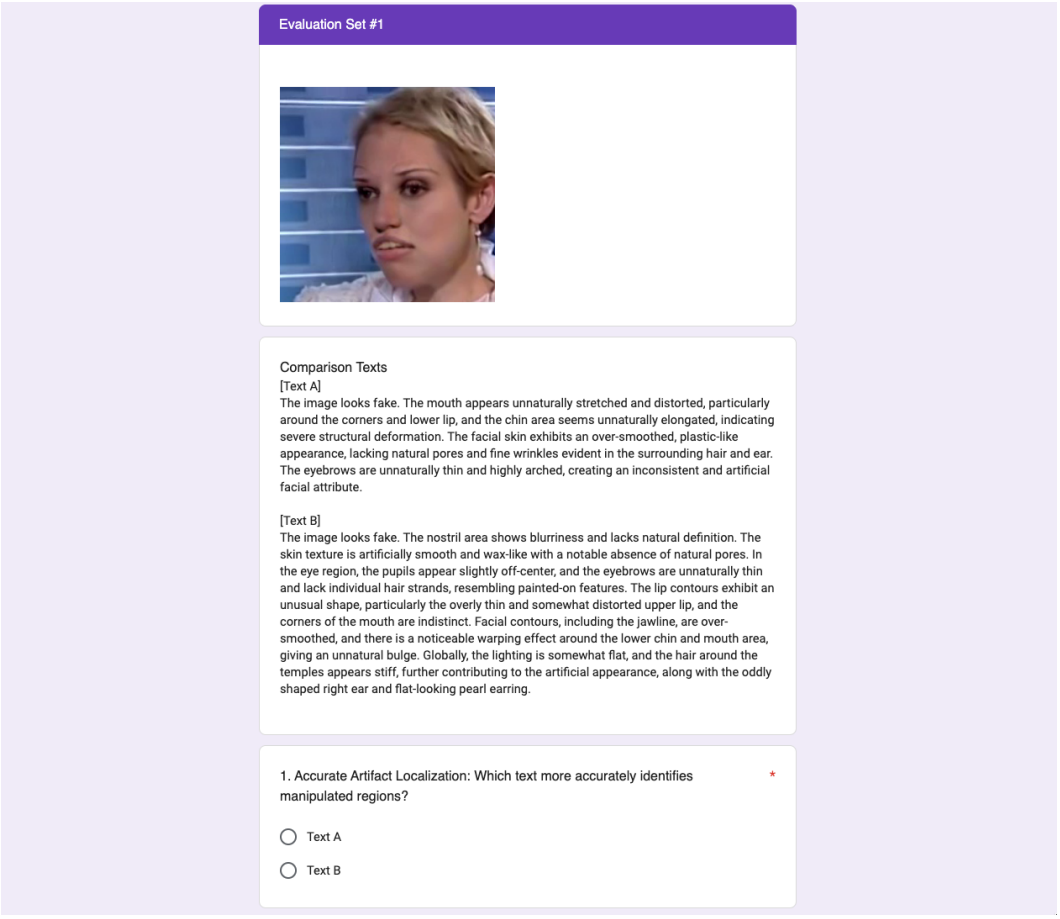


Figure 14: **User interface of the human evaluation (Survey 2).** For each forged image, participants compare two anonymized captions, one from ReForensics and one standard caption, presented in random order, under four evaluation criteria. Each participant completes 20 such sets.

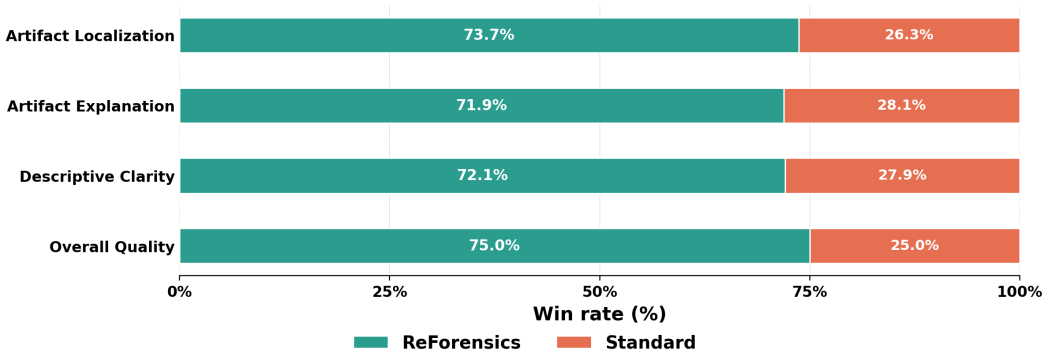


Figure 15: **Human evaluation of ReForensics captions (Survey 2).** Pairwise preference between ReForensics captions and standard captions. ReForensics captions are preferred in more than 71% of cases across all four criteria.

Annotation Quality Assessment (Score Evaluation)



Image A
(Reference)



Image B
(Target)

Context:
You are analyzing a Source-Target Image Pair to evaluate forensic captions.
Image A (Source/Real): The authentic original: { *Image A* }
Image B (Target/Deepfake): Generated from Image A: { *Image B* }
Text 1: A caption describing artifacts in Image B: { *Text 1* }
Text 2: A caption describing artifacts in Image B: { *Text 2* }
Goal: Evaluate the Reliability and Technical Depth of the text.

Evaluation Protocol
1. Artifact Localization Hallucination
- Criteria: Evaluates whether the identified artifacts are **genuine new errors** introduced during generation, rather than existing features from the source. Score each text based on this criterion. ***0 (FAIL) or 1 (PASS)***

2. Visual-Logical Consistency
- Criteria: Assesses whether the textual description of physical defects **aligns with the actual pixel data** in the target image. Score each text based on this criterion. ***0 (FAIL) or 1 (PASS)***

3. Descriptive Ambiguity & Technical Precision
- Criteria: Evaluates the clarity of the forensic evidence by rewarding the use of domain-specific terminology over vague or subjective language. Score each text based on this criterion. ***0 (FAIL) or 1 (PASS)***

4. Comparative Reliability (Relative Assessment)
- Criteria: A holistic comparison between two captions to determine the **ultimate winner** based on accuracy, technical depth, and conciseness.
- **0 (Loser):** The caption is disqualified due to critical errors identified in preceding stages or exhibits lower reliability because of excessive verbosity used to mask a lack of technical precision.
- **1 (Winner):** The caption provides a high-density, expert-level analysis by prioritizing precise accuracy. It successfully pinpoints the causal mechanisms of the artifact (the how and why) while avoiding superficial descriptions or unverified shotgun claims.

Output Format
You must return the result strictly in the following JSON format. Do not include any other text. No markdown blocks.

```
{ "detailed_step_evaluation": {
    "step_1_source_check": {
      "text_1": { "score": 0, "reason": "Pass/Fail explanation based on Image A." },
      "text_2": { "score": 0, "reason": "Pass/Fail explanation based on Image A." }
    },
    "step_2_visual_grounding": {
      "text_1": { "score": 0, "reason": "Pass/Fail explanation based on Image B." },
      "text_2": { "score": 0, "reason": "Pass/Fail explanation based on Image B." }
    },
    "step_3_specificity": {
      "text_1": { "score": 0, "reason": "Specific/Vague explanation. Did it get bonus points?" },
      "text_2": { "score": 0, "reason": "Specific/Vague explanation. Did it get bonus points?" }
    } },
  "final_verdict_process": {
    "comparison_reasoning": "Compare Text 1 and Text 2. State if one was penalized for over-claiming/shotgun approach.",
    "winner": "text_0"
  },
  "final_score_assignment": {
    "text_1_final_score": 0,
    "text_2_final_score": 0
  } }
```

Figure 16: **Prompt for scoring the reliability of forensic reasoning.** This prompt is designed to evaluate and score the reliability of two forensic reasoning texts, explicitly denoted as "Text 1" and "Text 2". The assessment benchmarks the reports based on four criteria: (1) Artifact Localization Hallucination, (2) Visual-Logical Consistency, (3) Descriptive Ambiguity & Technical Precision, and (4) Comparative Reliability.

H ADDITIONAL QUALITATIVE RESULTS

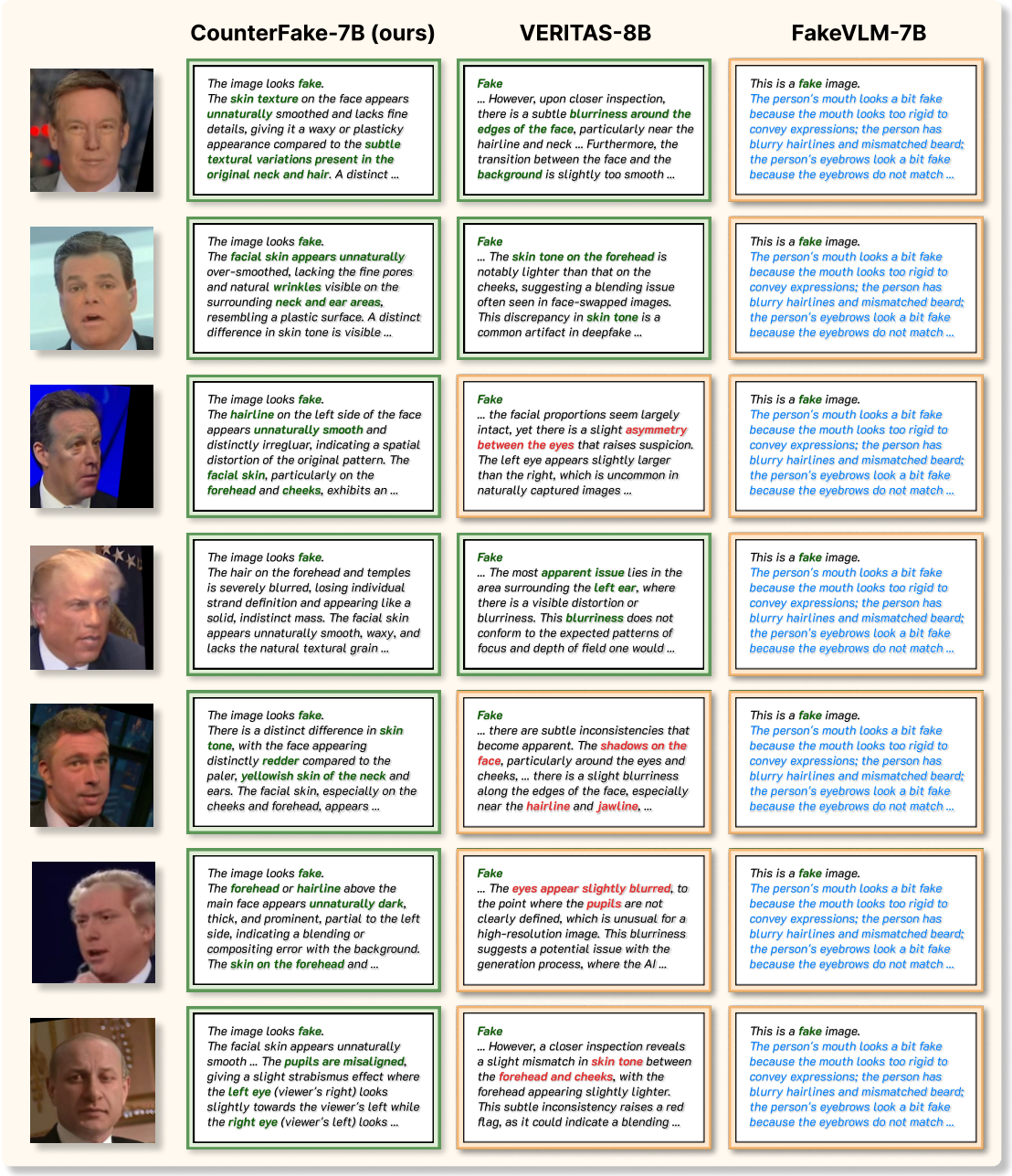


Figure 17: **Qualitative comparison of reasoning capabilities.** We compare rationales generated by CounterFake, VERITAS, and FakeVLM (from left to right) on representative fake samples. Green indicates a correct prediction with a faithful rationale, while Orange denotes a correct prediction supported by an incorrect rationale.