FORGE: Forensic Reasoning with Grounded Evidence

Rohit Kundu[1,2], **Shan Jia**[2], **Vishal Mohanty**[2],
Athula Balachandran[2], **Amit K. Roy-Chowdhury**[1]
[1]University of California, Riverside; [2]Google LLC
{rohit.kundu@email, amitrc@ece}.ucr.edu; {rohitkun, shanjia, vishalmohanty, athula}@google.com

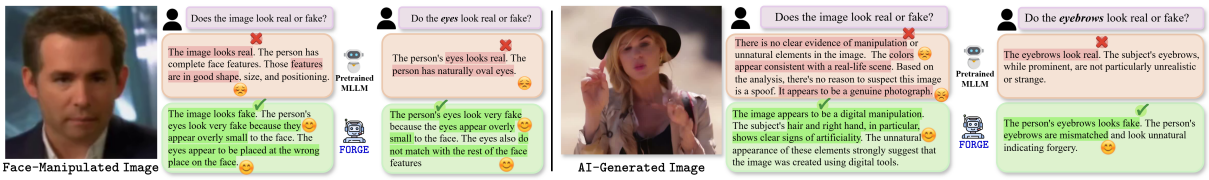


Figure 1: **Illustration of Examples**: Comparison of pretrained PaliGemma2 (Steiner et al., 2024) multimodal large language model's (MLLM) DeepFake reasoning against our proposed FORGE framework on a face-manipulated image (left) and a fully AI-generated image (right). The pretrained PaliGemma2 fails to detect subtle inconsistencies in the manipulated images, resulting in incorrect explanations. In contrast, FORGE provides accurate predictions along with detailed, fine-grained reasoning for the detected manipulations.

Abstract

Forensic deepfake analysis demands more than binary classification: investigators need region-grounded natural language explanations they can verify against the image. Multimodal large language models (MLLMs) are a natural fit, but pretrained MLLMs fail systematically, producing globally coherent text that misses the small localized cues defining manipulations. We argue this is an inductive bias problem rather than a capacity issue: the image-text contrastive objective training MLLM visual encoders optimizes for whole-image semantic summaries, not patch-level forensic detail. The same mismatch explains why prior deepfake reasoning methods target either face manipulation or fully AI-generated content, never both.

We propose FORGE, which addresses the mismatch by routing a second visual stream into the language model from a Vision-Only Model (VOM) trained on dense patch prediction rather than image-text alignment. The MLLM's native encoder and the VOM operate on a shared patch grid, which lets us interleave their tokens with preserved spatial correspondence; we show this beats naive concatenation. A two-stage adapter training protocol (generic image-caption alignment, then joint task-specific optimization) prevents the localized stream from overfitting to training-domain manipulations. Across face-manipulated and fully synthetic content, FORGE produces region-referential explanations answering fine-grained attribute queries ("Does the eyes/nose/mouth look real or fake?") and substantially outperforms in-domain baselines on cross-domain evaluations; region-specific evaluation and human studies confirm explanation faithfulness.

1 Introduction

Generative models for images and video have become both more accessible and more capable (Labs, 2024; Research, 2023; Wang et al., 2023a; Zheng et al., 2024), producing manipulated content that humans struggle to identify (Li et al., 2019; Thies et al., 2019). The forensic response to this has been a long line of detection methods that classify content as real or fake (Cheng et al., 2024; Kundu et al., 2025). But in deployment settings, a binary label is rarely sufficient. An investigator presented with a "fake" verdict needs to know which regions of the image triggered the decision and what about those regions looks manipulated, in natural language they can verify against the image itself. Naturally, we ask: *can multimodal large language models (MLLMs), optimized to describe images in broad strokes, be redirected to identify the small localized cues that define forensic manipulations?*

Recent work has explored using MLLMs for DeepFake reasoning, but the literature has fractured along the content type axis. One body of work targets face manipulation, where a real face

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 43745–43767
October 24-29, 2026 ©2026 Association for Computational Linguistics

image is altered by face swap, expression transfer, or attribute editing (Zhang et al., 2024; Chen et al., 2024b; Sun et al., 2025a; Guo et al., 2025). A separate body targets fully AI-generated content from text-to-image, text-to-video, or image-to-video models (Song et al., 2024; Yang et al., 2025). The two lines have proceeded essentially independently, with different datasets, different architectural choices, and different evaluation protocols. To our knowledge, no prior MLLM-based DeepFake reasoning method jointly addresses both content categories within a single framework.

This bifurcation is not coincidental. Pretrained MLLMs, evaluated on either task, fail in a specific and consistent way: they produce globally fluent descriptions that miss the small localized cues that distinguish manipulated from authentic content (Figure 1). We argue this is not a capacity issue, since using larger MLLMs for the problem does not resolve it (Jia et al., 2024), but an inductive bias issue. The image-text contrastive objective used to train MLLM visual encoders (CLIP (Radford et al., 2021), SigLIP (Zhai et al., 2023), and variants) optimizes representations for whole-image semantic summarization, the wrong target for forensic reasoning where the evidence lives in localized patches (Dahou et al., 2025; Tong et al., 2024). Naively fine-tuning the MLLM on forensic data papers over this on in-domain test sets but does not transfer across content categories or across manipulation types within a category.

We propose FORGE, which addresses the inductive bias mismatch directly. Rather than relying on the MLLM's contrastive encoder alone, we route a second visual stream into the language model from a Vision-Only Model (VOM) trained on dense patch prediction rather than image-text alignment, an encoder whose features are inherently localized by construction. The two streams are fused through a token-interleaved combination at a shared patch grid, which preserves spatial correspondence in a way that naive concatenation does not (Table 2). A lightweight adapter aligns the VOM stream to the LLM's embedding space, trained in two stages: first on generic image-caption data to establish general visual-linguistic alignment, then jointly with the LLM on forensic data. The first stage prevents the adapter from overfitting to training-domain manipulation patterns; ablating it collapses cross-domain performance (Table 3). Because the VOM stream operates at patch resolution and is decoupled from face-cropping preprocessing, the same

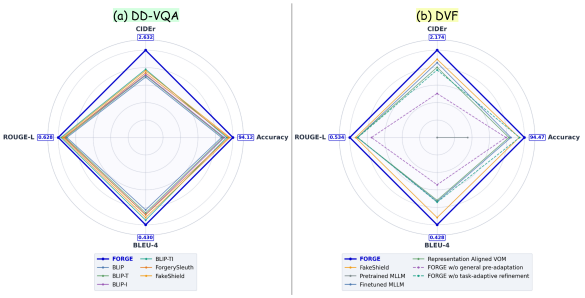


Figure 2: Comparison of FORGE and state-of-the-art DeepFake reasoning methods on two benchmarks. (a) DD-VQA (Zhang et al., 2024) (face-manipulated) and (b) DVF (Song et al., 2024) (fully AI-generated): Axis markings indicate the actual scores for Accuracy and scaled scores for BLEU-4, and ROUGE-L, and CIDEr. FORGE consistently outperforms competitors across all explanation and detection metrics, despite their specialization for face manipulation in (a) and significantly outperforms pretrained and simple finetuned MLLMs in (b), showing that our strategic grounding is essential for reasoning.

framework handles face-manipulation cues (asymmetric eyes, blending artifacts) and fully-synthetic cues (implausible textures, scene-level inconsistency) without separate architectures or training pipelines.

The contributions of this work are as follows.

- We diagnose a systematic failure mode of pretrained MLLMs on DeepFake reasoning and trace it to the image-text contrastive training of their visual encoders, an inductive bias problem rather than a capacity problem, not addressable by scale or naive fine-tuning alone.
- We propose a dual-stream visual pathway that pairs the MLLM with a Vision-Only Model (VOM) trained on dense patch prediction, using token-interleaved fusion at a shared 14×14 patch resolution to preserve spatial correspondence; we show that this spatial alignment, not the combination of encoders alone, is what enables region-referential generation.
- We show that two-stage adapter training (general visual-linguistic alignment, followed by joint task-specific optimization) is necessary to preserve cross-domain generalization, ablating each stage to demonstrate the specific failure modes it prevents.
- FORGE is the first MLLM-based DeepFake reasoning framework to address both face-manipulated and fully synthetic content within a single architecture. We validate it through

region-specific evaluation (eyes, nose, mouth for face manipulation; lighting, shadows, textures for fully synthetic content), human studies of explanation faithfulness, and cross-domain detection accuracy that outperforms methods trained in-domain on the evaluation datasets.

2 Related Work

Detection without Reasoning: Early DeepFake detection methods (Li, 2018; Yang et al., 2019) struggled with cross-dataset generalization, a problem subsequently addressed through identity-disentangled features (Dong et al., 2023), frequency-domain analysis (Corvi et al., 2023a,b), and reconstruction-based scores (Wang et al., 2023b). UNITE (Kundu et al., 2025) is the only method explicitly designed for both face-manipulated and fully synthetic content, using a transformer with attention-diversity loss, but like its predecessors it produces only binary labels. A separate thread surfaces detector decisions through visual saliency: Grad-CAM analyses (Țânțaru et al., 2024; Xu et al., 2022) highlight regions of high gradient response, and identity-prototype methods (Aghasanli et al., 2023) localize manipulations through distances to known identities. Both are limited as forensic tools: heatmaps are spatially coarse and show where the model attended rather than what is wrong, and prototype methods generalize poorly outside training identities. None of these approaches produce the natural language explanations one can verify against the image.

Face Manipulation Reasoning: One body of work targets face-manipulation DeepFakes (face swap, expression transfer, attribute editing) and generates natural language explanations through fine-tuned MLLMs. Early attempts evaluated GPT-4V (Achiam et al., 2023) and Gemini-1.0 (Anil et al., 2023) in zero-shot settings (Jia et al., 2024; Shi et al., 2024), finding that off-the-shelf MLLMs underperform task-specialized detectors. Subsequent methods finetune on annotated DeepFake data: DD-VQA (Zhang et al., 2024) finetunes BLIP (Li et al., 2022); X^2-DFD (Chen et al., 2024b) finetunes LLaVA (Liu et al., 2024); VLFFD (Sun et al., 2025a) uses visual-language prompts to guide detection; M2F2-Det (Guo et al., 2025) combines CLIP with a Vicuna-7b decoder; ForgerySleuth (Sun et al., 2025b) adds a trace encoder for MLLM-based manipulation analysis; FFAA (Huang et al., 2024) proposes a multi-answer decision system for open-world face forgery analysis; and Fake-in-Facext (Qin et al., 2025) introduces a fine-grained facial concept tree for attribute-level explanation. These methods share two limitations relevant to our setting: they detect and crop faces as a preprocessing step (assuming a front-view face is always visible), and they are evaluated only on face-manipulated content and cannot describe the scene-level cues that define fully synthetic images.

Synthetic Content Reasoning: A symmetric body of work targets fully AI-generated content from text-to-image (T2I), text-to-video (T2V), and image-to-video (I2V) models, where manipulations affect entire scenes rather than localized regions. MM-Det (Song et al., 2024) finetunes LLaVA (Liu et al., 2024) on diffusion-generated video with text annotations to generate explanations. HEIE (Yang et al., 2025) produces hierarchical implausibility heatmaps with occasional textual descriptions, committing primarily to spatial output. More recent work explores preference optimization with auxiliary visual experts for AI-generated image detection (Zhou et al., 2025) and two-stage scan-and-zoom strategies that locate suspicious regions before generating verdicts (Ji et al., 2025). Symmetrically to the face-manipulation methods, these systems are evaluated only on fully synthetic content and do not detect or explain face-manipulation cues that affect a small region of an otherwise authentic image.

Positioning of FORGE: The two bodies of work above have proceeded essentially independently, training on different datasets, using different architectures, and reporting different evaluation protocols. To the best of our knowledge, no prior MLLM-based DeepFake reasoning method addresses both content categories within a single framework. We argue in Section 1 that this bifurcation reflects the inductive bias mismatch of MLLM visual encoders rather than coincidence. Two recent methods come close to a unified treatment and warrant explicit positioning. FakeShield (Xu et al., 2025) covers Photoshop tampering, DeepFake attribute editing, and partial AIGC inpainting (a real image with a region redrawn by Stable Diffusion), but does not address fully synthetic content where the entire image is generated from scratch (T2I, T2V, I2V). IVY-FAKE (Jiang et al., 2025) unifies image and video AI-generated content detection with natural language explanations, but is not evaluated on face-manipulation benchmarks where

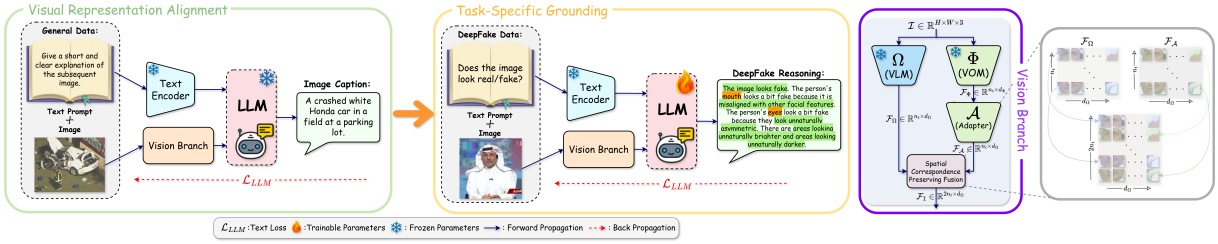


Figure 3: **Architecture of FORGE:** The framework routes two visual streams with complementary inductive biases into the language model. The contrastive encoder Ω (VLM) supplies global semantic features that are natively compatible with the LLM, while the vision-only model Φ (VOM) supplies dense patch features through a learnable adapter $\mathcal{A}$ that maps them into the LLM's embedding space. The two streams operate on a shared patch grid, which lets us combine them via spatial-correspondence-preserving fusion (right) without learned fusion layers. Training proceeds in two stages: visual representation alignment (left), where $\mathcal{A}$ is calibrated on general image-caption data with the LLM frozen, then task-specific grounding (center), where $\mathcal{A}$ and the LLM are jointly optimized on language-annotated deepfake data. The same architecture handles face-manipulated and fully AI-generated content.

a real face is altered. FORGE addresses the original bifurcation directly: face-manipulation cues that affect small regions of otherwise authentic images, and synthesis cues that affect entire scenes, within a single architecture trained on a single set of language-annotated data.

3 Proposed Method

3.1 Problem Setup

Unlike existing methods (Chen et al., 2024b; Zhang et al., 2024; Sun et al., 2025a; Guo et al., 2025) that focus solely on face-manipulated DeepFakes, our model provides a unified framework for explainability in both face-manipulated and fully AI-generated images. Given an image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ and a question $\mathcal{Q}$ related to DeepFake detection, the objective of our model is to generate a detailed answer (ground truth answer: $\mathcal{E}$) that determines whether the image is real or manipulated and provides an explanation for the decision. Our goal is to find the optimal parameters θ of our model f_θ that minimize the following objective function:

$$\arg\min_{\theta} \sum_{(\mathcal{I},\mathcal{Q},\mathcal{E})\in\mathcal{D}} \mathcal{L}(f_\theta(\mathcal{I},\mathcal{Q}),\mathcal{E}) \qquad (1)$$

where $\mathcal{D}$ is our training dataset of image-question-explanation triplets, and $\mathcal{L}$ is a loss function measuring discrepancy between generated and ground truth explanations. This formulation enables our model to learn accurate, detailed explanations for DeepFake images.

3.2 Complementary Visual Inductive Biases

As argued in Sec. 1, the failure mode of pretrained MLLMs on this task is an inductive bias mismatch: a visual encoder optimized for whole-image alignment with captions cannot reliably surface the patch-level cues that define face manipulations. We respond architecturally rather than by scaling. We leave the MLLM's native visual stream untouched and route a second stream into the language model from a VOM trained on dense patch prediction. The two streams are chosen for their structurally different training objectives, not as redundant feature extractors.

Global Context: The MLLM's native visual encoder (Ω) is trained with image-text contrastive loss, which optimizes representations for whole-image alignment with caption embeddings. This objective is well-suited to scene-level reasoning, the level at which fully synthetic content typically reveals itself: composition, lighting, object arrangement, and semantic plausibility. Ω partitions $\mathcal{I}$ into non-overlapping $p \times p$ patches and produces,

$$\mathcal{F}_\Omega = \Omega(\mathcal{I}^{H\times W\times 3}) \in \mathbb{R}^{n_t\times d_\Omega}, \qquad (2)$$

where $n_t = \frac{H\times W}{p^2}$. These n_t tokens carry the holistic scene signal that the LLM already consumes natively, so we route them into the language model directly without transformation.

Localized Context: Contrastive training compresses per-patch detail in service of whole-image alignment, which is what produces the systematic blindness to small localized cues documented in Figure 4: asymmetric eyes, unnaturally curved nose lines, faint blending boundaries. A model fluent on global semantics but blind on local patches will reliably miss the evidence that defines face-manipulation cues, even after task-specific finetuning (Table 2). We address this by routing a second

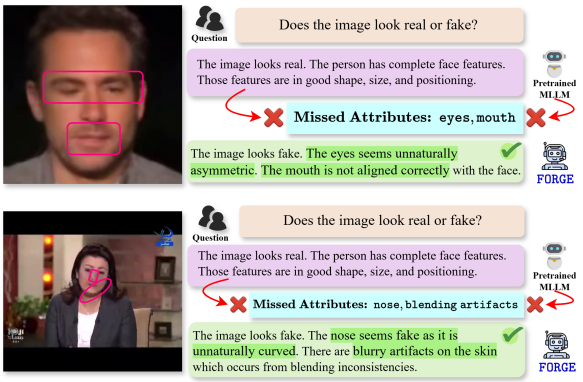


Figure 4: Pretrained MLLMs miss the small localized cues that define faceial DeepFakes. FORGE surfaces them by routing a dense-prediction stream into the language model alongside the contrastive stream.

stream into the language model from a VOM (Φ) trained on dense patch prediction. Self-supervised dense prediction optimizes for patch-level distinctiveness rather than caption alignment, preserving the localized signal that contrastive training discards. Φ also operates by partitioning input $\mathcal{I}$ into non-overlapping $p \times p$ patches and produces,

$$\mathcal{F}_\Phi = \Phi(\mathcal{I}^{H\times W\times 3}) \in \mathbb{R}^{n_t \times d_\Phi}. \tag{3}$$

These n_t patch tokens are language-unaligned by construction (Φ has never been trained against text), which motivates the cross-modal adapter introduced in Sec. 3.3. Critically, both encoders operate on the same $p \times p$ patch grid: the i-th token in $\mathcal{F}_\Omega$ and the i-th token in $\mathcal{F}_\Phi$ describe the same spatial region. This shared grid is what makes spatial-correspondence-preserving fusion possible.

3.3 Spatial-Correspondence-Preserving Fusion

Fusing the two streams faces two structural challenges that distinguish this setting from generic multimodal fusion. First, the streams live on different manifolds by design: $\mathcal{F}_\Omega$ is aligned with the LLM's caption-trained embedding space, while $\mathcal{F}_\Phi$ is shaped by dense patch prediction and language-unaligned. Second, they describe the same image at the same spatial resolution: the i-th token of each stream corresponds to the same image patch (Sec. 3.2). A fusion design that ignores either fact will perform suboptimally. We address each below.

Adapter: Cross-Manifold Projection. $\mathcal{F}_\Omega$ already lives in the LLM's input space and is consumed directly. $\mathcal{F}_\Phi$ requires a projection, but generic dimension matching is not sufficient: the adapter must move features from a manifold optimized for patch-level distinctiveness into a manifold optimized for whole-image alignment with captions, without collapsing the patch-level distinctions that motivated routing the VOM into the model in the first place. We instantiate $\mathcal{A}$ as a lightweight learnable map,

$$\mathcal{F}_\mathcal{A} = \mathcal{A}(\mathcal{F}_\Phi) \in \mathbb{R}^{n_t \times d_\Omega}, \tag{4}$$

and observe that how $\mathcal{A}$ is trained matters more than how it is parameterized: the two-stage protocol in Sec. 3.4 is what gives $\mathcal{A}$ its locality-preserving behavior, and without it the same architecture collapses (Table 3).

Token-Aligned Interleaving: With $\mathcal{F}_\Omega$ and $\mathcal{F}_\mathcal{A}$ in a common embedding space, the remaining question is how to combine them. The obvious choice is concatenation along the token dimension,

$$\mathcal{F}_\mathrm{C} = \mathcal{F}_\Omega \| \mathcal{F}_\mathcal{A} \in \mathbb{R}^{2n_t \times d_\Omega}. \tag{5}$$

Concatenation is lossless, but it discards the structural fact from Sec. 3.2 that $\mathcal{F}_\Omega[i]$ and $\mathcal{F}_\mathcal{A}[i]$ describe the same spatial region. Under concatenation, this correspondence has to be recovered by the LLM through learned attention across a token distance of n_t positions. A design that honors the shared grid by construction avoids this. We interleave the streams token by token,

$$\mathcal{F}_\mathrm{I}^k = \begin{cases} \mathcal{F}_\Omega^j, & k = 2j-1 \\ \mathcal{F}_\mathcal{A}^j, & k = 2j \end{cases} \tag{6}$$
$$j \in \{1, 2, \ldots, n_t\}, k \in \{1, 2, \ldots, d_\Omega\}$$

The contrastive and dense-prediction views of every spatial region now appear as adjacent tokens. The LLM no longer has to learn which global token corresponds to which local token; the answer is given by position. This design adds no learnable parameters, so the gain over concatenation in Table 2 cannot be attributed to additional capacity: it is the inductive bias of the fusion operator itself.

3.4 Disentangled Adapter Training

The adapter $\mathcal{A}$ has two purposes: align $\mathcal{F}_\Phi$ with the LLM's input space (a general visual-linguistic problem), and preserve patch-level distinctiveness through the projection (a forensic problem, since it is this distinctiveness that surfaces face-manipulation cues). Optimizing both jointly on DeepFake data alone couples them: the adapter learns whichever projection minimizes loss on the training distribution, which is typically a shortcut

that ties locality to a small set of manipulation patterns and does not generalize. We decouple the two goals by training in two stages. Stage 1 establishes general visual-linguistic alignment before any DeepFake data is seen, and Stage 2 specializes the now-aligned adapter and LLM jointly on the forensic task. Table 3 confirms that removing either stage, or collapsing them into a single joint run, degrades performance significantly.

Visual Representation Alignment: We train $\mathcal{A}$ on the LCS-558K image-caption corpus (Liu et al., 2024) with the LLM frozen, optimizing $\mathcal{L}_{\text{LLM}}$ (defined below) over standard caption generation. This stage sees no DeepFake data, no forensic vocabulary, and no region-level supervision. Its only objective is to teach $\mathcal{A}$ how to map $\mathcal{F}_\Phi$ into a space the language model can consume, using whatever visual concepts a generic caption corpus contains. The result is an adapter aligned to language at the manifold level, before it is asked to surface any task-specific cue. Skipping this stage (*No general pre-adaptation* in Table 3) collapses cross-domain performance, because $\mathcal{A}$ has no general visual-linguistic alignment to fall back on when the task distribution shifts.

Task-Specific Grounding: With $\mathcal{A}$ now language-aligned, we unfreeze the LLM and jointly optimize $\mathcal{A}$ and the LLM on language-annotated DeepFake data. $\mathcal{A}$ remains trainable rather than frozen: freezing it (*No task-adaptive refinement* in Table 3) leaves the adapter unable to specialize its projection for forensic attributes, and forensic precision drops accordingly. The joint update lets the adapter sharpen its locality-preservation in directions that matter for the task while the LLM learns the corresponding linguistic mapping. Collapsing Stage 1 and Stage 2 into a single joint run on DeepFake data (*Joint training only*) achieves a middle outcome: the adapter has no foundational alignment to anchor it, so its task specialization is tied to manipulation patterns in the training set and transfers poorly across content categories.

Loss Objective: Both stages share the same autoregressive language-modeling loss. For each example, the question $\mathcal{Q}$ and reference answer $\mathcal{E}$ are tokenized as

$$\mathcal{T}_{\text{prefix}} = \tau(\mathcal{Q}), \quad \mathcal{T}_{\text{separator}} = \tau(\text{“}\backslash n\text{”}), \quad (7)$$
$$\mathcal{T}_{\text{suffix}} = \tau(\mathcal{E}), \mathcal{T} = \mathcal{T}_{\text{prefix}} \| \mathcal{T}_{\text{separator}} \| \mathcal{T}_{\text{suffix}}.$$

Cross-entropy is computed only over answer tokens, enforced by a loss mask $\mathcal{M}_{\text{loss}} = [0]^{|\mathcal{T}_{\text{prefix}}|+|\mathcal{T}_{\text{separator}}|} \| [1]^{|\mathcal{T}_{\text{suffix}}|}$. $\mathcal{T}_{\text{prefix}}$ attend bidirectionally and $\mathcal{T}_{\text{suffix}}$ attend causally to support autoregressive generation. The objective is, thus,

$$\mathcal{L}_{\text{LLM}} = -\frac{\sum_{i=1}^{|\mathcal{T}|} \mathcal{M}_{\text{loss}}^i \cdot \log P(\mathcal{T}^i)}{\sum_{i=1}^{|\mathcal{T}|} \mathcal{M}_{\text{loss}}^i}, \quad (8)$$

where $P(\cdot)$ is the model's predicted distribution over tokens. At inference, only $\mathcal{T}_{\text{prefix}}$ is provided and the model generates responses autoregressively.

4 Experiments

Training & Evaluation Protocol: FORGE is trained on DD-VQA (Zhang et al., 2024) (face-manipulated FF++ (Rossler et al., 2019) with language annotations) and DVF (Song et al., 2024) (fully synthetic, language-annotated). Detection performance is also evaluated cross-domain on CelebDF (Li et al., 2020) and DF40 (Yan et al., 2024b) (face-manipulated) and DeMamba (Chen et al., 2024a) (fully synthetic), which lack language annotations. Dataset details are in §A1; implementation details (hardware/software) in §A2.

For each example, we report (i) an *LLM-as-a-judge accuracy* that uses Gemini-1.0 (Anil et al., 2023) to check whether the model's verdict (real vs. fake) agrees with the reference; and (ii) standard generation metrics BLEU-3, BLEU-4, ROUGE-L, and CIDEr against the reference explanation, following (Chen et al., 2024b; Zhang et al., 2024). These two families of metrics are necessary but not sufficient. The binary judge captures whether conclusions agree, not whether the supporting reasoning is faithful, and n-gram metrics correlate weakly with human judgments of explanation quality (Novikova et al., 2017). We therefore complement them with a dedicated explanation-quality evaluation: a three-annotator human study on 300 stratified samples scoring faithfulness, specificity, and forensic helpfulness; a rubric-based GPT-4o judge calibrated against the human ratings (Pearson $r \geq 0.76$ on every dimension) and applied at full scale; and a faithfulness analysis that quantifies hallucinated region and attribute claims. Full protocol, instructions, and detailed results are in §A3.

State-of-the-art Comparison: Table 1 compares FORGE against baselines on DD-VQA (face-manipulated) and DVF (fully synthetic). MLLM-based reasoning baselines exist only on the face-manipulated side (BLIP-TI, X^2-DFD,

Table 1: **SOTA Comparison**. The competitor methods are specifically designed for either face-manipulation or synthetic data reasoning. However our FORGE framework is designed for generalized reasoning on both face-manipulated and synthetic content, and still outperforms in both categories. **Best** and second-best performances are marked. *-marked methods were run using official implementation and model weights.

Dataset	Method	Accuracy	AUC	BLEU-4	ROUGE-L	CIDEr
Face Manipulated Data						
DD-VQA (Zhang et al., 2024)	XceptionNet-BLIP-TI (Zhang et al., 2024)	89.25%	92.24%	-	-	-
	HifiNet (Guo et al., 2023)	89.16%	92.10%	-	-	-
	HifiNet-BLIP-TI (Zhang et al., 2024)	91.25%	95.14%	-	-	-
	RECCE (Cao et al., 2022)	91.03%	95.02%	-	-	-
	RECCE-BLIP (Zhang et al., 2024)	89.22%	93.71%	-	-	-
	RECCE-BLIP-TI (Zhang et al., 2024)	92.08%	95.36%	-	-	-
	BLIP (Li et al., 2022)	81.68%	-	0.3569	0.5664	1.8177
	BLIP-T (Zhang et al., 2024)	83.65%	-	0.3714	0.5774	1.8715
	BLIP-I (Zhang et al., 2024)	84.87%	-	0.3800	0.5882	1.8931
	BLIP-TI (Zhang et al., 2024)	87.49%	-	0.4075	0.6085	2.0567
	$\mathcal{X}^2$-DFD (Chen et al., 2024b)	-	-	0.2030	0.1550	0.0270
	ForgerySleuth* (Sun et al., 2025b)	89.42%	93.18%	0.3812	0.5854	1.9486
	FakeShield* (Xu et al., 2025)	90.85%	93.91%	0.3947	0.5982	2.0143
	FORGE	**94.12%**	**95.39%**	**0.4304**	**0.6285**	**2.6321**
Synthetic Data						
DVF (Song et al., 2024)	CNNDet (Wang et al., 2020)	-	78.20%	-	-	-
	DIRE (Wang et al., 2023b)	-	62.10%	-	-	-
	(Cozzolino et al., 2024)	-	67.00%	-	-	-
	UNI-FD (Ojha et al., 2023)	-	74.10%	-	-	-
	F3Net (Qian et al., 2020)	-	81.30%	-	-	-
	ViViT (Arnab et al., 2021)	-	79.10%	-	-	-
	TALL (Xu et al., 2023)	-	69.50%	-	-	-
	TS2-Net (Liu et al., 2022)	-	72.10%	-	-	-
	DE-FAKE (Sha et al., 2023)	-	72.10%	-	-	-
	HifiNet (Guo et al., 2023)	-	84.30%	-	-	-
	DVF (Song et al., 2024)	-	92.00%	-	-	-
	FakeShield* (Xu et al., 2025)	88.74%	89.83%	0.3924	0.4985	1.9482
	FORGE	**94.47%**	**95.82%**	**0.4279**	**0.5338**	**2.1744**

Table 2: **Architectural ablations on FORGE's vision pathway.** From a pretrained MLLM (capacity baseline), through single-stream and dual-stream variants, to spatial-correspondence-preserving fusion. **Best** and second-best performances are marked.

Dataset	Configuration	Accuracy ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	CIDEr ↑
DD-VQA (Zhang et al., 2024)	Pretrained MLLM	11.50%	0.0000	0.0000	0.0218	0.2610
	MLLM only ($\mathcal{F}_\Omega$)	88.19%	0.4012	0.3768	0.5633	2.1478
	VOM only ($\mathcal{F}_\mathcal{A}$)	90.45%	0.4107	0.4143	0.5843	2.0816
	FORGE w/ concat. ($\mathcal{F}_\mathrm{C}$)	93.56%	0.4497	0.4286	0.6015	2.4193
	FORGE ($\mathcal{F}_\mathrm{I}$)	**94.12%**	**0.4649**	**0.4304**	**0.6285**	**2.6321**
DVF (Song et al., 2024)	Pretrained MLLM	33.07%	0.0000	0.0000	0.0034	0.0001
	MLLM only ($\mathcal{F}_\Omega$)	79.88%	0.3142	0.3839	0.4877	1.8629
	VOM only ($\mathcal{F}_\mathcal{A}$)	77.67%	0.3077	0.3786	0.4873	1.7465
	FORGE w/ concat. ($\mathcal{F}_\mathrm{C}$)	92.81%	0.3614	0.4261	0.5234	2.1605
	FORGE ($\mathcal{F}_\mathrm{I}$)	**94.47%**	**0.3786**	**0.4279**	**0.5338**	**2.1744**

Table 3: **Ablation on Adapter Configurations.**

Dataset	Adapter Settings	Accuracy ↑	BLEU-3 ↑	BLEU-4 ↑	ROUGE-L ↑	CIDEr ↑
Face-Manipulated Data						
DD-VQA (Zhang et al., 2024)	No general pre-adaptation	79.13%	0.3017	0.2975	0.5113	1.5812
	No task-adaptive refinement	86.56% (↑7.43%)	0.3964 (↑0.0947)	0.3897 (↑0.0922)	0.5712 (↑0.0599)	2.0447 (↑0.4635)
	Joint training only	90.371% (↑3.81%)	0.4107 (↑0.0143)	0.4066 (↑0.0169)	0.6015 (↑0.0303)	2.3186 (↑0.2739)
	FORGE	**94.12%** (↑3.75%)	**0.4649** (↑0.0542)	**0.4304** (↑0.0238)	**0.6285** (↑0.0270)	**2.6321** (↑0.3135)
Synthetic Data						
DVF (Song et al., 2024)	No general pre-adaptation	74.39%	0.2317	0.2815	0.4014	1.0961
	No task-adaptive refinement	89.23% (↑14.84%)	0.3173 (↑0.0856)	0.3681 (↑0.0866)	0.4846 (↑0.0832)	1.6820 (↑0.5859)
	Joint training only	92.11% (↑2.88%)	0.3350 (↑0.0177)	0.3875 (↑0.0194)	0.4972 (↑0.0126)	1.8988 (↑0.2168)
	FORGE	**94.47%** (↑2.36%)	**0.3786** (↑0.0436)	**0.4279** (↑0.0404)	**0.5338** (↑0.0366)	**2.1744** (↑0.2756)

ForgerySleuth), reflecting the literature bifurcation; on the synthetic side, our reasoning-metric competitors are detection-only methods. The single baseline that operates on both regimes is FakeShield (Xu et al., 2025), which scores reasonably on DD-VQA (90.85% accuracy, 2.01 CIDEr) but lags on DVF (88.74% accuracy, 89.83% AUC, 1.95 CIDEr), illustrating that near-unified treatments still leave a coverage gap. FORGE leads on every metric on both datasets: +3.27% accuracy and +0.62 CIDEr over FakeShield on DD-VQA, and +5.73% accuracy, +5.99% AUC, and +0.23 CIDEr on DVF. Cross-domain detection results on CelebDF, DF40, and DeMamba (in §A4) show FORGE outperforming competitors evaluated in-domain despite being itself evaluated purely cross-domain.

Architectural Ablations: Table 2 walks five configurations from the off-the-shelf MLLM up to the full FORGE system, each falsifying a successively stronger "simpler explanation" for the gains. The off-the-shelf MLLM scores 11.50% accuracy on DD-VQA and 33.07% on DVF, well below the 50% binary baseline and consistent with the zero-shot failures reported in (Jia et al., 2024): scale alone does not solve the task. Finetuning the same MLLM on DeepFake reasoning data with no second stream (*naive grounding*, the $\mathcal{F}_\Omega$-only row) lifts performance to 88.19% / 79.88%; adding the VOM stream is what closes the remaining gap to the full system, contributing +5.9 / +14.6 accuracy points on DD-VQA / DVF. The single-stream VOM baseline ($\mathcal{F}_\mathcal{A}$-only) shows the opposite asymmetry: stronger on DD-VQA (90.45%, where face manipulations are localized patches) and weaker on DVF (77.67%, where scene-level cues matter), confirming that neither inductive bias is sufficient alone. Combining the two streams via concatenation ($\mathcal{F}_\mathrm{C}$) closes most of the gap to 93.56% / 92.81%, but token-aligned interleaving ($\mathcal{F}_\mathrm{I}$, ours) extracts a further 0.6-1.7 points on the same data, the same encoders, and the same training pipeline. Since interleaving adds zero learnable parameters, this final gain cannot be attributed to capacity; it is the spatial alignment of the fusion operator itself, as argued in Sec. 3.3.

Adapter Ablation: Table 3 tests the two-stage training protocol against three alternatives: skipping the foundational alignment stage (adapter trained only on DeepFake data; *No general pre-adaptation*), skipping the task-specific stage (adapter pretrained on LCS-558K but frozen during downstream training; *No task-adaptive refinement*), and collapsing the two stages into one (adapter and LLM jointly optimized on DeepFake data from scratch; *Joint training only*). Each maps to a specific prediction from Sec. 3.4, and each underperforms in the way that prediction implies. Removing the foundational stage costs the most: 79.13% / 74.39% on DD-VQA / DVF, the largest gap to the full system, because the adapter has no general visual-linguistic alignment to fall back on when the task distribution shifts. Freezing the adapter after the foundational stage reaches 86.56% / 89.23%;

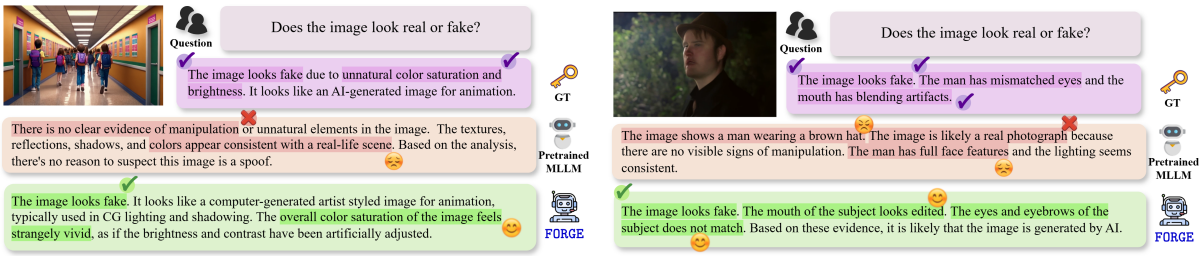


Figure 5: Examples demonstrating the DeepFake relevance of the responses generated by FORGE compared to a pretrained MLLM (PaliGemma2). FORGE not only enhances accuracy on DeepFake detection, but also improves relevance to the task providing finegrained reasoning. Additional examples in §A6

Table 4: **Faithfulness and hallucination analysis.** RHR: Region Hallucination Rate (fraction of explanations referencing an authentic region as manipulated). AHR: Attribute Hallucination Rate (fraction of attribute claims unsupported by ground-truth attribute labels). FP: Forensic Precision (fraction of attribute claims on fake images that align with annotated manipulation cues). FORGE cuts RHR by more than 2× over the strongest baseline on both datasets.

Dataset	Method	RHR ↓	AHR ↓	FP ↑
DD-VQA (Zhang et al., 2024)	Pretrained PaliGemma2 (Steiner et al., 2024)	64.3%	71.2%	18.4%
	ForgerySleuth (Sun et al., 2025b)	39.7%	43.8%	50.1%
	FakeShield (Xu et al., 2025)	36.4%	40.2%	53.6%
	BLIP-TI (Zhang et al., 2024)	41.2%	45.6%	48.7%
	Grounded PaliGemma2 (no VOM)	28.9%	31.5%	67.8%
	FORGE	**12.7%**	**14.3%**	**84.6%**
DVF (Song et al., 2024)	Pretrained PaliGemma2 (Steiner et al., 2024)	58.7%	66.4%	22.1%
	FakeShield (Xu et al., 2025)	38.6%	42.1%	51.8%
	MM-Det (Song et al., 2024)	35.2%	38.9%	55.4%
	Grounded PaliGemma2 (no VOM)	30.1%	33.8%	64.2%
	FORGE	**15.4%**	**17.2%**	**81.3%**

Table 5: **VOM ablation:** Replacing DINOv2 with MAE (ViT-L/14) isolates the contribution of dense patch prediction as a training paradigm from the specific choice of encoder. Both dense-prediction encoders substantially exceed the contrastive-only baseline (row-1), and DINOv2's small residual advantage is consistent with feature-space over pixel-space reconstruction.

VOM (Φ)	DD-VQA (Zhang et al., 2024)			DVF (Song et al., 2024)			Params
	BLEU-4 ↑	ROUGE-L ↑	CIDEr ↑	BLEU-4 ↑	ROUGE-L ↑	CIDEr ↑	
$\mathcal{F}_\Omega$ only (no VOM)	0.3768	0.5633	2.1478	0.3839	0.4877	1.8629	–
MAE ViT-L/14	0.4143	0.6104	2.5391	0.4142	0.5187	2.1105	304M
DINOv2 ViT-L/14 (ours)	**0.4304**	**0.6285**	**2.6321**	**0.4279**	**0.5338**	**2.1744**	304M

the alignment is in place but the adapter cannot sharpen its projection for forensic attributes. Collapsing the two stages reaches 90.37% / 92.11%, the closest to FORGE and the most informative comparison: the adapter has access to all the same data but couples the alignment and specialization jobs, and the remaining 2-4% gap is what disentangling them recovers.

VOM Ablation: The architectural argument in Sec. 3.2 attributes the second stream's contribution to dense patch prediction as a training paradigm, not to DINOv2 specifically. Table 5 tests this claim by replacing DINOv2 with MAE (ViT-L/14, ImageNet-pretrained), keeping every other component of FORGE identical. MAE shares DINOv2's patch-level supervision paradigm (masked patch reconstruction) but reconstructs in pixel space rather than feature space and lacks DINOv2's multi-crop student-teacher stabilization. MAE recovers most of DINOv2's gain over the contrastive-only $\mathcal{F}_\Omega$ baseline (within 2-4% relative on every metric on both datasets), and both dense-prediction encoders substantially exceed $\mathcal{F}_\Omega$ alone. This confirms that the gain comes from the dense patch prediction paradigm itself: any encoder whose training objective preserves per-patch distinctiveness supplies the localized signal the contrastive encoder discards. DINOv2's small residual advantage over MAE is consistent with feature-space reconstruction preserving semantic patch distinctions that pixel-space reconstruction does not, but the effect is second-order compared to the paradigm switch from contrastive to dense-prediction training.

Table 6: **Fusion strategy ablation:** Alternative parameterized fusion mechanisms compared to token-aligned interleaving on DD-VQA, holding all other components of FORGE fixed. Interleaving matches or outperforms every parameterized alternative while adding zero learnable parameters.

Fusion Strategy	Added Params	Accuracy ↑	BLEU-4 ↑	CIDEr ↑
Q-Former	∼12M	93.10%	0.4190	2.5100
Cross-attention	∼4M	93.80%	0.4270	2.5800
Gated fusion	∼1M	93.70%	0.4250	2.5900
Interleaving (ours)	**0**	**94.12%**	**0.4304**	**2.6321**

Fusion Strategy Ablation: Table 6 compares token-aligned interleaving against three parameterized alternatives on DD-VQA: cross-attention (learned key/value projections between streams), Q-Former (compressive cross-modal projection), and gated fusion (per-token learned gates). All three underperform interleaving despite adding capacity, and Q-Former, the fusion strategy behind BLIP-2 and InstructBLIP, underperforms most. This isolates the contribution to the fusion operator's inductive bias rather than to any learned combination: preserving spatial correspondence by construction beats learning to recover it.

Region-Referential Reasoning: Detection accuracy can be high for the wrong reasons, so we

directly evaluate whether FORGE's explanations are faithful to the image and specific enough to be forensically useful. Three annotators (calibrated with a 90-minute domain expert session) rate 300 stratified samples on faithfulness, specificity, and forensic helpfulness (1-5 Likert; Krippendorff's $\alpha \geq 0.68$; full protocol in §A3). Table A3 reports the results. FORGE leads every baseline on every dimension on both datasets, with the largest gains on specificity and forensic helpfulness (+0.77 and +0.65 over the grounded PaliGemma2 ablation on DD-VQA, with comparable margins on DVF), the dimensions most directly tied to region-referential reasoning. We complement perceived quality with an objective check: Table 4 reports the rate at which generated explanations hallucinate manipulated regions (RHR) or unsupported attribute claims (AHR), and the precision of attribute claims on manipulated images (FP). FORGE reaches RHR of 12.7% / 15.4% on DD-VQA / DVF, less than half the next-best baseline (28.9% / 30.1%), and lifts forensic precision past 80% on both datasets. The grounded PaliGemma2 ablation, which removes only the VOM stream, retains the higher hallucination rates, confirming that detection-finetuning alone does not produce faithful reasoning. A per-region breakdown across face attributes (eyes, nose, mouth, etc.) and scene aspects (lighting, shadows, textures, etc.) is in §A7.

Qualitative Results: Figure 5 shows side-by-side reasoning from a pretrained MLLM (PaliGemma2) and FORGE on face-manipulated and fully synthetic inputs. The pretrained MLLM produces fluent but generic descriptions that miss the manipulation: a face-swap with mismatched eyes is described as having "complete face features... in good shape, size, and positioning," and an AI-generated scene is read as a real photograph. FORGE identifies the specific region and the specific attribute that defines the manipulation, naming asymmetric eyes, blending artifacts on the skin, mismatched mouth alignment, and unnaturally vivid color saturation as appropriate. The failure mode of the pretrained MLLM and the success mode of FORGE on the same inputs are the qualitative analogue of the quantitative gap in Tables 4 and A3. Additional examples are in §A6.

5 Conclusion

We presented FORGE, an MLLM-based framework for DeepFake reasoning that addresses an inductive bias mismatch in pretrained MLLM visual encoders: contrastive training optimizes for whole-image alignment with captions, not the patch-level cues that define forensic manipulations. FORGE routes a second stream into the language model from a vision-only model trained on dense patch prediction, fuses the two streams at a shared patch grid via spatial-correspondence-preserving interleaving, and trains the adapter in two disentangled stages. A single architecture handles both face-manipulated and fully AI-generated content, outperforming specialized baselines on each, with region-referential explanations whose faithfulness is confirmed by human evaluation and a hallucination analysis. The prior literature's bifurcation between these two content categories reflects this inductive bias problem rather than an inherent difference between the regimes; resolving the mismatch lets a single framework cover both.

Limitations

FORGE focuses on image-level DeepFake reasoning and does not model temporal anomalies in videos. Extending the framework to video is a natural direction: the dual-stream design and interleaving fusion generalize to per-frame or per-clip inputs without architectural change, and the primary constraint is the current absence of publicly available benchmarks with frame-level or temporal textual annotations for DeepFake forensics. As such benchmarks emerge, FORGE's architecture is well-positioned to extend directly.

Our evaluation is anchored on the two publicly available datasets that provide language-annotated DeepFake reasoning supervision at scale, DD-VQA (face manipulation) and DVF (fully synthetic content). Together these span the two dominant manipulation regimes we set out to unify, and our results demonstrate strong performance across both. Manipulation categories outside these two regimes, such as partial inpainting or targeted attribute editing, are addressed by separate lines of work (e.g., FakeShield) and are complementary to our contribution rather than in scope; extending region-referential reasoning to these regimes is a promising next step as language-annotated benchmarks in those categories become available.

Acknowledgements: This work was supported in part by funding from YouTube (Google LLC).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agil Aghasanli, Dmitry Kangin, and Plamen Angelov. 2023. Interpretable-through-prototypes deepfake detection for diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 467–474.

Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. 2024. Getting vit in shape: Scaling laws for compute-optimal model design. *Advances in Neural Information Processing Systems*, 36.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, and 1 others. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, and 1 others. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. 2022. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and 1 others. 2024a. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*.

Yize Chen, Zhiyuan Yan, Siwei Lyu, and Baoyuan Wu. 2024b. $\mathcal{X}^2$-dfd: A framework for e$\mathcal{X}$plainable and e$\mathcal{X}$tendable deepfake detection. *arXiv preprint arXiv:2410.06126*.

Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. 2024. Can we leave deepfake data behind in training deepfake detector? *arXiv preprint arXiv:2408.17052*.

Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. 2024. Exploiting style latent flows for generalizing deepfake video detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1133–1143.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2023a. 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*, pages 973–982.

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

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

Yasser Dahou, Ngoc Dung Huynh, Phuc H Le-Khac, Wamiq Reyaz Para, Ankit Singh, and Sanath Narayan. 2025. Vision-language models can't see the obvious. *arXiv preprint arXiv:2507.04741*.

Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. 2023. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3994–4004.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595.

Deepfakes GitHub. 2017. https://github.com/deepfakes/faceswap.

FaceSwap GitHub. 2016. https://github.com/MarekKowalski/FaceSwap/.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. 2021. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM*

international conference on multimedia, pages 3473–3481.

Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. 2023. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. 2025. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 105–116.

Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. 2023. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4490–4499.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Yikun Ji, Yan Hong, Bowen Deng, Huijia Zhu, Weiqiang Wang, Liqing Zhang, Jianfu Zhang, and 1 others. 2025. Zoom-in to sort ai-generated images out. *arXiv preprint arXiv:2510.04225*.

Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. 2024. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4324–4333.

Changjiang Jiang, Wenhui Dong, Zhonghao Zhang, Chenyang Si, Fengchang Yu, Wei Peng, Xinbin Yuan, Yifei Bi, Ming Zhao, Zian Zhou, and 1 others. 2025. Ivy-fake: A unified explainable framework and benchmark for image and video aigc detection. *arXiv preprint arXiv:2506.00979*.

Haojun Jiang, Jianke Zhang, Rui Huang, Chunjiang Ge, Zanlin Ni, Jiwen Lu, Jie Zhou, Shiji Song, and Gao Huang. 2022. Cross-modal adapter for text-video retrieval. *arXiv preprint arXiv:2211.09623*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, and Amit K Roy-Chowdhury. 2025. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Blackforest Labs. 2024. Flux 1.1. https://blackforestlabs.ai/.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2019. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*.

Y Li. 2018. Exposing deepfake videos by detecting face warping artif acts. *arXiv preprint arXiv:1811.00656*.

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

Yuzhen Lin, Wentang Song, Bin Li, Yuezun Li, Jiangqun Ni, Han Chen, and Qiushi Li. 2024. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In *European Conference on Computer Vision*, pages 104–122. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. 2022. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European conference on computer vision*, pages 319–335. Springer.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1906–1919.

Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. 2022. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12–21.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2241–2252.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, and 1 others. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer.

Lixiong Qin, Yang Zhang, Mei Wang, Jiani Hu, Weihong Deng, and Weiran Xu. 2025. Fake-in-facext: Towards fine-grained explainable deepfake analysis. *arXiv preprint arXiv:2510.20531*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Runway Research. 2023. Text driven video generation. https://research.runwayml.com/gen2.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*.

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. 2023. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3418–3432.

Yichen Shi, Yuhao Gao, Yingxin Lai, Hongyang Wang, Jun Feng, Lei He, Jun Wan, Changsheng Chen, Zitong Yu, and Xiaochun Cao. 2024. Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. *arXiv preprint arXiv:2402.04178*.

Kaede Shiohara and Toshihiko Yamasaki. 2022. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18720–18729.

Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. 2024. On learning multi-modal forgery representation for diffusion generated video detection. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, and 1 others. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. 2025a. Towards general visual-linguistic face forgery detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19576–19586.

Zhihao Sun, Haoran Jiang, Haoran Chen, Yixin Cao, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. 2025b. Forgerysleuth: Empowering multimodal large language models for image manipulation detection. *Advances in Neural Information Processing Systems*, 38:153136–153175.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. 2024. 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*, pages 28130–28139.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide

shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023a. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. 2020. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. 2023b. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455.

Ying Xu, Kiran Raja, and Marius Pedersen. 2022. Supervised contrastive learning for generalizable and explainable deepfakes detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 379–389.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. 2023. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22658–22668.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. 2025. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *International Conference on Learning Representations*, volume 2025, pages 31186–31216.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. 2024a. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8994.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Li Yuan, Chengjie Wang, Shouhong Ding, and 1 others. 2024b. Df40: Toward next-generation deepfake detection. *NeurIPS 2024*.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. 2023. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22412–22423.

Fan Yang, Ru Zhen, Jianing Wang, Yanhao Zhang, Haoxiang Chen, Haonan Lu, Sicheng Zhao, and Guiguang Ding. 2025. Heie: Mllm-based hierarchical explainable aigc image implausibility evaluator. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3856–3866.

Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. 2024. Common sense reasoning for deepfake detection. In *European Conference on Computer Vision*, pages 399–415. Springer.

Cairong Zhao, Chutian Wang, Guosheng Hu, Haonan Chen, Chun Liu, and Jinhui Tang. 2023. Istvt: interpretable spatial-temporal video transformer for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 18:1335–1348.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. 2024. Open-sora: Democratizing efficient video production for all.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. 2025. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18746–18758.

Dragoș-Constantin Țânțaru, Elisabeta Oneață, and Dan Oneață. 2024. Weakly-supervised deepfake localization in diffusion-generated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6258–6268.

This additional material provides extended details and analyses supporting the main paper. Sec. A1 describes the datasets and the exact splits used. Sec. A2 gives implementation details (hardware, software, and model configurations). Sec. A3 presents the full explanation-quality evaluation summarized in the main paper: the human-study protocol and annotator instructions, the calibrated rubric judge, and the faithfulness and hallucination analysis. Sec. A4 reports cross-domain detection results on CelebDF, DF40, and DeMamba. Sec. A5 examines robustness to image compression, and Sec. A6 provides additional qualitative examples.

A1 Details of the Datasets

The detailed descriptions for all the datasets used for training and evaluation of our FORGE framework are as follows:

- **DD-VQA** (Zhang et al., 2024)**:** This dataset was created by annotating one frame per video from the FaceForensics++ (FF++) (Rossler et al., 2019) dataset. The original FF++ (Rossler et al., 2019) datasets consists of 1000 real videos, and DeepFake videos generated by four methods: Face2Face (Thies et al., 2016), FaceSwap (GitHub, 2016), DeepFakes (GitHub, 2017), and NeuralTextures (Thies et al., 2019), totalling 4000 videos.

  For each image, there may be several question-answer pairs, and not every video has a language annotation available. Considering each image-language pair as a single sample, the total number of samples for each resolution of FF++ (Rossler et al., 2019) are: (1) raw: 14298 (2) c23: 14298 (3) c40: 19803. This disparity depending on the resolutions is because, the fake videos created using the NeuralTextures (Thies et al., 2019) method is only available for the c40 split, and not the others.

- **CelebDF** (Li et al., 2020)**:** This dataset contains DeepFake videos featuring celebrity faces, designed to reduce noticeable artifacts and enhance realism compared to earlier datasets. It incorporates variations in lighting, backgrounds, and facial movements for diversity. CelebDF includes 590 real videos and 5639 DeepFake videos. No language annotations exist for the CelebDF (Li et al., 2020) dataset.

- **DF40** (Yan et al., 2024b)**:** The DF40 (Yan et al., 2024b) dataset comprises over 1 million DeepFake samples, generated using 40 distinct face manipulation techniques, including face-swapping, face-reenactment, entire face synthesis, and face editing. It includes both image- and video-level data, offering a diverse and large-scale benchmark for DeepFake detection. The dataset emphasizes forgery realism and diversity, simulating real-world scenarios with aligned data domains from FF++ and CDF. This comprehensive dataset facilitates robust evaluations across various manipulation techniques, supporting advancements in generalizable DeepFake detection.

- **DVF** (Song et al., 2024)**:** The Diffusion Video Forensics (DVF) dataset is a benchmark designed for detecting forgery in diffusion-generated videos. It comprises 2750 real videos and 3938 fake videos generated using eight diffusion-based methods, including text-to-video (T2V) and image-to-video (I2V) models. The dataset includes diverse resolutions and durations, with fake videos sourced from both open-sourced and closed-sourced methods. Notably, language annotations are available only for videos generated by the Stable Video Diffusion model (Blattmann et al., 2023). DVF serves as a comprehensive benchmark for evaluating forgery detection methods across a wide range of video manipulation techniques.

- **DeMamba** (Chen et al., 2024a)**:** This recently introduced dataset is created using various text-to-video (T2V) and image-to-video (I2V) models, with the training and validation sets generated by distinct models. The training set comprises 1.2 million real videos and 1.08 million fake videos, while the validation set includes 10000 real videos and 9588 fake videos. For our cross-dataset analyses in the main paper, we utilize only the validation set.

- **LCS-558K** (Liu et al., 2024)**:** The LCS-558K dataset (used for our adapter training in Stage-1) is a subset of the LAION (Schuhmann et al., 2022), CC (Changpinyo et al., 2021), and SBU (Ordonez et al., 2011) datasets, filtered to ensure balanced concept coverage across visual and textual modalities. It consists of $558,000$ image-text pairs and is primarily

used for pretraining multimodal models. The dataset includes BLIP-synthesized captions to enhance alignment between visual and textual features, making it suitable for tasks such as visual instruction tuning and multimodal understanding.

For training FORGE, we utilized the designated training split from DD-VQA (Zhang et al., 2024) and allocated 80% of the DVF (Song et al., 2024) dataset for training, as an explicit split for the language annotated data was not provided by the original authors. The remaining 20% of DVF was reserved for evaluation. All other datasets were employed exclusively in cross-dataset evaluation. Specifically, for DeMamba (Chen et al., 2024a), we adopted the official evaluation split released with the dataset. For other evaluation benchmarks lacking standardized splits, we followed established protocols in the literature (Xu et al., 2023; Chen et al., 2024b; Sun et al., 2025a), randomly selecting 20% of the data for evaluation. This approach ensures consistency with prior work and allows for reliable cross-dataset comparison in our experiments.

All datasets and pretrained models used in this work are publicly released for research purposes, and we use them only in that capacity. FORGE, trained on these research-only resources, is intended exclusively for research on DeepFake forensics.

We do not collect any new data. All datasets used are established, publicly released research benchmarks, and we use them as distributed under their original terms. The face-centric datasets necessarily contain images of identifiable individuals, because facial identity is intrinsic to the face-manipulation detection task and cannot be anonymized without destroying the forensic signal the task depends on; collection, consent, and release of these subjects were handled by the original dataset creators. We add no new identifying metadata, do not attempt to re-identify or link individuals across datasets, and use the data solely for research on DeepFake forensics. We are not aware of offensive content beyond the manipulated/synthetic imagery inherent to DeepFake benchmarks, which we use only as detection targets.

The 80/20 split on DVF is stratified by generator (all eight generators appear in both partitions) and by real/fake label, so both training and evaluation sets have identical class balance. There is no generator-level leakage in the sense of train-only generators appearing at test time. There is also no near-duplicate leakage: DVF videos are synthesized independently by different generators from different prompts and are not near-duplicates of one another by construction. Train and test videos are disjoint at the file level. We followed the split protocol used by (Song et al., 2024) and by follow-up work.

A2 Implementation Details

FORGE is trained using Stochastic Gradient Descent with an initial learning rate of 0.0001, scheduled to decay following a cosine annealing schedule. The training process is conducted with a batch size of 64 over 5 epochs for both representation alignment of Φ and task-targeted forensic grounding. Ω is the PaliGemma2 (Beyer et al., 2024) MLLM's visual encoder, SigLIP-So400m/14 (Alabdulmohsin et al., 2024), Φ is the DINOv2 (Oquab et al., 2023) vision-only foundation model and τ is the Gemma2 (Rivière et al., 2024) tokenizer in our case. In Eqs. 2 and 3, image resolutions $H = W = 448$ and patch size $p = 14$, yielding $n_t = 1024$. The feature dimensions $d_\Omega = 2304$ and $d_\Phi = 2048$ are a property of the foundation models. Adapter $\mathcal{A}$ is composed of one MLP layer following popular choices in the literature (Gao et al., 2024; Jiang et al., 2022). FORGE has a total of 3.3B total parameters (2.6B trainable), with each stage trained for 5 epochs. The framework is implemented in JAX and executed on 8 TPUv3 chips, ensuring efficient large-scale training.

Computational cost: Inference on a single A100 (448×448 input):

- FORGE: 720ms/image, 12.4GB peak memory
- Grounded PaliGemma2 ($\mathcal{F}_\Omega$ only): 480ms, 8.7GB
- FakeShield (LLaVA-7B backbone): 1100ms, 15.2GB
- ForgerySleuth (LLaVA-7B backbone): 1250ms, 17.1GB

The added VOM stream contributes 240ms and 3.7GB, a modest overhead relative to accuracy and faithfulness gains. FORGE is faster and lighter than both MLLM-based mask-generating baselines because it does not run a separate mask decoder. We will include this comparison in the appendix.

Table A1: **Pretrained vs. Grounded MLLM:** Performance comparison between pretrained and finetuned PaliGemma2 (Steiner et al., 2024). Although grounding PaliGemma2 improves performance FORGE still performs better with its local context harvesting (as shown in Table 2).

Dataset	Grounded	Accuracy ↑	BLEU-3 ↑	BLUE_4 ↑	ROUGE-L ↑	CIDEr ↑
Face Manipulated Data						
DD-VQA (Zhang et al., 2024)	✗	11.50%	0.0000	0.0000	0.0218	0.2610
	✓	88.19% (+76.69%)	0.4012 (+0.4012)	0.3768 (+0.3768)	0.5633 (+0.5415)	2.1478 (+1.8868)
Synthetic Data						
DVF (Song et al., 2024)	✗	33.07%	0.0000	0.0000	0.0034	0.0001
	✓	79.88% (+46.81%)	0.3142 (+0.3142)	0.3839 (+0.3839)	0.4877 (+0.4843)	1.8629 (+1.8628)

A3 Explanation Quality Evaluation

The detection-oriented metrics in Section 4 establish that FORGE reaches and improves the conclusion a forensic system is asked to reach. They do not establish that it reaches the conclusion *for the right reasons*. For a system whose deployment value is the explanation itself, this is the central question. Surface-level n-gram metrics (BLEU, ROUGE, CIDEr) are known to correlate weakly with human judgments of generated explanations (Novikova et al., 2017), and our binary LLM-as-judge accuracy in Table 1 reports only whether predicted and ground-truth conclusions agree, not whether the supporting reasoning is sound.

We therefore complement those metrics with three studies aimed directly at explanation quality: (i) a multi-annotator human evaluation along forensically meaningful dimensions (§A3.1); (ii) a rubric-based LLM-as-judge calibrated against the human ratings, which we use to scale evaluation to the full test sets (§A3.2); and (iii) a faithfulness and hallucination analysis that quantifies how often generated explanations reference regions or attributes that are not actually manipulated (§A3.3). Together, these analyses provide direct evidence that the detection accuracy reported in Section 4 is grounded in faithful, region-referential reasoning rather than coincidence.

A3.1 Human Evaluation Protocol

Data: We sample 300 image-explanation pairs: 150 from DD-VQA (Zhang et al., 2024) and 150 from DVF (Song et al., 2024). Sampling is stratified across (a) the real/fake label and (b) manipulation subtype, so each subtype contributes proportionally and no system is advantaged by an easy slice. For every image we collect candidate explanations from six systems: pretrained PaliGemma2 (Steiner et al., 2024); FakeShield (Xu et al., 2025), the strongest publicly comparable mask-based explanation baseline whose released weights handle both face-manipulated and fully synthetic content; ForgerySleuth (Sun et al., 2025b), the closest competing image-manipulation MLLM with public weights (DD-VQA only, as it is not trained for fully synthetic video); BLIP-TI (Zhang et al., 2024) (DD-VQA only) or MM-Det (Song et al., 2024) (DVF only) as the strongest dataset-native baseline; the grounded PaliGemma2 ablation (Table A1, i.e. FORGE minus the VOM stream); and FORGE. FakeShield and ForgerySleuth are run from their official public checkpoints with the original inference configurations; we use only their textual explanation outputs, ignoring their mask outputs for the purposes of comparing language quality.

Annotators: Three graduate-student annotators with prior research experience in vision-language modeling rated each explanation. Prior to annotation, all three completed a 90-minute calibration session led by a domain expert in image forensics, covering common deepfake artifact families (blending boundaries, iris and pupil asymmetries, texture statistics, lighting/shadow inconsistencies, geometric implausibilities). Annotators saw the image, the question, and the candidate explanations in randomized order with system identifiers removed. Annotators were not told which explanation was generated by which system, including not being told that one was the ground-truth reference. Each annotator independently rated all 300×6 explanations across three dimensions, then ratings were averaged across annotators per (image, system, dimension) cell.

Annotator Recruitment and Independence. The three annotators for the human evaluation were graduate students in the broader lab whose research interests span vision-language modeling but who had no involvement in this specific project. None was an author of the paper, participated in the design of the method, reviewed any draft prior to annotation, supervised or funded any component of the work, or had prior exposure to the training data, the model outputs, or the specific claims being tested. During annotation, system identities were removed from the interface, the presentation order of candidate explanations was randomized per image, and annotators were not told which explanation was the ground-truth reference or how many systems were being compared. The 90-minute calibration session covered common DeepFake artifact families (blending boundaries, eye asymmetries,

texture statistics, lighting and shadow inconsistencies, geometric implausibilities) in general terms; it made no reference to our system, training data, architectural choices, or claims under test, and all calibration examples were drawn from publicly available datasets outside our evaluation splits. Recruitment from within the institution was a procedural choice that simplified approvals, scheduling, and the domain-expert calibration session; it does not reflect any prior familiarity of the annotators with our work. The inter-annotator agreement reported in Table A3 (Krippendorff's $\alpha \geq 0.68$ on every dimension) supports the interpretation that ratings were driven by the annotation criteria rather than by project-specific priors, consistent with recruitment practice in comparable NLP evaluation studies.

Rating dimensions: Each explanation is rated on a 1-5 Likert scale along three forensically motivated axes:

- **Faithfulness:** does the explanation describe cues that are actually present in the image? Penalizes hallucinated artifacts and fabricated regions.
- **Specificity:** does the explanation name concrete regions and attributes (*"the iris is asymmetric"*) rather than make generic claims (*"the image looks fake"*)?
- **Forensic helpfulness:** if presented to an investigator, would the explanation support a useful follow-up examination of the image?

These three axes are deliberately separated: a generic explanation can be faithful but unhelpful, and a confidently specific one can be unfaithful. Detection-oriented metrics conflate all three.

Annotator instructions (verbatim): The instruction sheet given to each annotator is reproduced below to enable reproduction.

You will see a series of images, each paired with a question ("Does the X look real or fake?") and several candidate explanations produced by different AI systems. The system identities are hidden and the order is randomized for every image. For each explanation, assign three independent scores on a 1-5 scale:

Faithfulness: *Does every cue named in the explanation correspond to something visible in the image?* 1 = multiple confidently asserted artifacts that are not present (severe hallucination); 2 = at least one major unsupported claim; 3 = mostly accurate but contains at least one minor unsupported claim; 4 = all major claims are verifiable; minor wording could be tighter; 5 = every cue cited is independently verifiable in the image.

Specificity: *Does the explanation name concrete regions and attributes, or does it speak in generalities?* 1 = entirely generic ("the image looks off"); 2 = vague references to "the face" or "the lighting" without an attribute; 3 = names a region but not the specific attribute that is anomalous; 4 = names both a region and an attribute (e.g. "the eyes are asymmetric"); 5 = names both a specific region and a specific localized attribute (e.g. "the iris of the left eye is asymmetric relative to the right").

Forensic helpfulness: *If you were an investigator deciding where to examine the image next, how useful is the explanation as a starting point?* 1 = no actionable information; 2 = some direction but too vague to act on; 3 = points to a region but not what to look for; 4 = points to a region and an artifact, but other regions might also be relevant; 5 = an investigator could immediately verify the cues claimed and would not waste time on irrelevant regions.

Score the three dimensions independently: a generic-but-correct explanation can score high on faithfulness and low on specificity. Do not let the question wording bias your reading; rate the explanation against the image, not against the question.

Table A2: Inter-annotator agreement (Krippendorff's α, ordinal) across three annotators on the 300-sample human evaluation set.

Dataset	Faith.	Spec.	Foren. Help.
DD-VQA (Zhang et al., 2024)	0.74	0.71	0.69
DVF (Song et al., 2024)	0.73	0.70	0.68
Pooled	**0.74**	**0.71**	**0.68**

Agreement: Inter-annotator agreement is reported using Krippendorff's α for ordinal data (Krippendorff, 2011), computed per dimension and per dataset; agreement is substantial across the board (Table A2). Faithfulness has the highest agreement, consistent with prior findings that hallucinations are easier for annotators to flag than to score severity for (Maynez et al., 2020).

Results: Mean per-system ratings are reported in Table A3. FORGE improves over every baseline on every dimension and on both datasets. The improvement over the grounded PaliGemma2 ablation is sharpest on *specificity* ($+0.77$ on DD-VQA, $+0.88$ on DVF) and *forensic helpfulness* ($+0.65$ on DD-VQA, $+0.78$ on DVF), which is the prediction made by our central architectural argument: the dense-patch VOM stream supplies the localized evidence that an image-text contrastive encoder systematically smooths over. Notably, FORGE also outperforms the mask-based explanation baselines FakeShield and ForgerySleuth on every dimension, despite producing region-referential text rather than spatial masks: text grounding to the right regions is competitive with, and on these axes preferred over, mask grounding of comparable specificity. Pretrained PaliGemma2 scores well below 2 on every axis, confirming that capacity alone does not solve the task.

Table A3: Human evaluation (1-5 Likert, mean of three annotators) on 300 stratified samples. FORGE leads on every dimension and dataset; gains are largest on specificity and forensic helpfulness, the dimensions most directly tied to region-referential explanation. Inter-annotator agreement is reported in Table A2.

Dataset	Method	Faithfulness ↑	Specificity ↑	Forensic Helpfulness ↑	Average ↑
DD-VQA (Zhang et al., 2024)	Pretrained PaliGemma2 (Steiner et al., 2024)	1.82	1.54	1.71	1.69
	ForgerySleuth (Sun et al., 2025b)	2.96	2.75	2.85	2.85
	BLIP-TI (Zhang et al., 2024)	3.04	2.87	2.91	2.94
	FakeShield (Xu et al., 2025)	3.08	2.84	2.95	2.96
	Grounded PaliGemma2 (no VOM)	3.58	3.41	3.47	3.49
	FORGE	**4.31**	**4.18**	**4.12**	**4.20**
DVF (Song et al., 2024)	Pretrained PaliGemma2 (Steiner et al., 2024)	1.94	1.68	1.83	1.82
	FakeShield (Xu et al., 2025)	2.97	2.61	2.79	2.79
	MM-Det (Song et al., 2024)	3.21	2.83	2.99	3.01
	Grounded PaliGemma2 (no VOM)	3.42	3.18	3.31	3.30
	FORGE	**4.24**	**4.06**	**4.09**	**4.13**

A.3.2 Rubric-Based LLM-as-Judge

Human evaluation is the gold standard, but does not scale to the full DD-VQA and DVF test sets. We therefore introduce a rubric-based LLM-as-judge and validate it against the 300-sample human study before applying it more broadly.

Judge design: We prompt GPT-4o (Hurst et al., 2024) with the image, the ground-truth explanation, the candidate explanation, and a rubric that defines the same three dimensions used by the human annotators. The judge produces a 1-5 score per dimension along with a one-sentence rationale. Crucially, the judge model is different from the model used in our detection-accuracy metric (Gemini-1.0), so judge predictions for explanation quality are not collusive with our detection metric.

Rubric prompt (verbatim): The exact prompt issued to GPT-4o per (image, candidate) pair, with placeholders shown in angle brackets, is reproduced below.

```
You are an expert evaluator of deepfake forensic explanations.
You are given:
- An image.
- A question that was asked about the image.
- A ground-truth explanation that describes whether the image is authentic or manipulated, and which regions or attributes support that conclusion.
- A candidate explanation produced by an AI system.
Score the candidate explanation on a 1-5 scale on each of the three dimensions below, INDEPENDENTLY. Use the full range. Do not let a high score on one dimension inflate another.
1. FAITHFULNESS: Does every cue the candidate names correspond to something present in the image? Penalize hallucinated artifacts.
1 = multiple major hallucinations.
2 = at least one major unsupported claim.
3 = mostly accurate with at least one minor unsupported claim.
4 = all major claims are verifiable.
5 = every cue cited is independently verifiable in the image.
2. SPECIFICITY: Does the candidate name concrete regions AND specific attributes, or speak in generalities?
1 = entirely generic.
2 = vague region reference, no attribute.
3 = region named, attribute missing.
4 = region and attribute named.
5 = specific sub-region (e.g., left iris) and specific attribute named.
3. FORENSIC HELPFULNESS: Could an investigator use this explanation to choose where to look next in the image?
1 = no actionable content.
2 = vague direction.
3 = region pointed to but unclear what to look for.
4 = region + artifact, but other regions might also be relevant.
5 = an investigator could immediately verify the specific cues.
Output STRICT JSON of the form:
{"faithfulness": <int 1-5>, "specificity": <int 1-5>, "forensic_helpfulness": <int 1-5>, "rationale": "<one sentence>"}
Question: <question>
Ground-truth explanation: <gt_explanation>
Candidate explanation: <candidate_explanation>
```

We use temperature 0 for determinism and validate that every response is parseable JSON (failures are re-prompted up to 3 times; in practice $< 0.5\%$ of calls required a retry).

Calibration against humans: On the 300-sample human-annotated subset, the rubric judge correlates strongly with the averaged human rating: Pearson $r = 0.83/0.79/0.76$ and Spearman $\rho = 0.81/0.77/0.74$ on faithfulness, specificity, and forensic helpfulness respectively (Table A4). All correlations are above the $r \geq 0.6$ threshold typically required for an LLM judge to be considered a usable proxy for human ratings (Zheng et al., 2023). As a system-level check, the judge preserves the human-induced ranking of all six systems on both datasets.

Full-test-set results: Applying the calibrated rubric judge to the full DD-VQA and DVF test

Table A4: Calibration of the rubric-based GPT-4o judge against the mean of three human annotators on the 300-sample subset. All correlations exceed the 0.6 threshold for usable judge reliability.

Dimension	Pearson r	Spearman ρ
Faithfulness	0.83	0.81
Specificity	0.79	0.77
Forensic Helpfulness	0.76	0.74
Average	**0.81**	**0.79**

sets (Table A5) preserves the human-induced ordering up to within-noise swaps of the two closest baselines (FakeShield and BLIP-TI, separated by 0.02 in the human study and 0.13 in the rubric judge, both well within the judge's calibration noise). FORGE leads on every dimension on both datasets, and the gap to grounded PaliGemma2 widens slightly compared to the 300-sample subset, suggesting that the human-set gap is not the result of favorable sampling.

Our detection-accuracy metric already uses Gemini-1.0 as the judge, so the two evaluation studies deliberately use different LLM families (Gemini for detection, GPT-4o for rubric quality). We additionally cross-validated the rubric judge with Gemini-2.5 and Claude-3.5-Sonnet on the same 300-sample calibration subset. All three judges preserve the same system-level ordering, with system-level Kendall's $\tau > 0.85$ across judges and Pearson $r > 0.72$ on each dimension. Beyond LLM judges, our faithfulness analysis (RHR, AHR, FP) uses no LLM in the scoring loop: it compares structured claim extractions against ground-truth region/attribute annotations. The faithfulness result is therefore judge-free and shows the same ordering.

A3.3 Faithfulness and Hallucination Analysis

The previous two studies measure perceived explanation quality. We now ask a sharper question: when a FORGE explanation references a region or attribute, is that reference *warranted by the image*? Both datasets include region- or attribute-level supervision that lets us compute this directly. DD-VQA annotates eyes, nose, mouth, eyebrows, and skin as authentic or manipulated. DVF annotates background, lighting, shadows, textures, reflections, object boundaries, and color consistency. We treat these annotations as a reference and compute three metrics:

- **Region Hallucination Rate (RHR↓):** fraction of generated explanations that reference at least one *authentic* region as manipulated.
- **Attribute Hallucination Rate (AHR↓):** fraction of distinct attribute-level claims (e.g. "asymmetric eyes," "inconsistent lighting") in the generated explanation that are not supported by the ground-truth attribute labels for that image.
- **Forensic Precision (FP↑):** of all attribute-level claims made by the model on *fake* images, the fraction that align with annotated manipulation cues.

Claim extraction: To map free-form explanations to (region, attribute, polarity) tuples we prompt GPT-4o with the structured extraction template reproduced below. We validate the extractor on a 200-sample subset against expert-parsed tuples, obtaining 96.4% precision and 94.1% recall; on disagreement, the dominant failure mode is the extractor splitting a single human-perceived claim across two tuples, which is not consequential for the aggregate metrics.

```
You are extracting structured forensic claims from a free-form deepfake explanation. Output a JSON list of claims; each claim is an object with the keys below.
- "region": one of {eyes, nose, mouth, eyebrows, skin, hair, ears, face_global, background, lighting, shadows, textures, reflections, object_boundaries, color, global}, or "none" if no region is named.
- "attribute": the specific anomaly mentioned (e.g., "asymmetric", "blending_artifact", "inconsistent_lighting", "unnatural_texture", "overexposed", "implausible_geometry"), or "none".
- "polarity": "manipulated" if the claim asserts manipulation in this region/attribute, "authentic" if it asserts authenticity, "unsure" if hedged (e.g., "may be slightly off").
Rules:
- One tuple per distinct claim. Do not invent claims not present in the text.
- If a region is referenced without a specific attribute, set "attribute" to "none" but keep the tuple.
- If the explanation contains no concrete claims, output [].
Example input:
"The eyes look asymmetric and the skin near the chin has blending artifacts, but the lighting is consistent."
Example output:
[{"region":"eyes","attribute":"asymmetric",
"polarity":"manipulated"},
{"region":"skin","attribute":"blending_artifact",
"polarity":"manipulated"},
```

Table A5: Rubric-based GPT-4o judge results over the full DD-VQA and DVF test sets. Ordering matches the 300-sample human study (Table A3); system gaps widen modestly at scale, indicating the human-set advantage of FORGE is not an artifact of sampling.

Dataset	Method	Faithfulness ↑	Specificity ↑	Forensic Helpfulness ↑	Average ↑
DD-VQA (Zhang et al., 2024)	Pretrained PaliGemma2 (Steiner et al., 2024)	1.93	1.61	1.84	1.79
	ForgerySleuth (Sun et al., 2025b)	2.91	2.69	2.81	2.80
	FakeShield (Xu et al., 2025)	3.02	2.78	2.91	2.90
	BLIP-TI (Zhang et al., 2024)	3.12	2.95	3.02	3.03
	Grounded PaliGemma2 (no VOM)	3.64	3.49	3.55	3.56
	FORGE	**4.36**	**4.24**	**4.18**	**4.26**
DVF (Song et al., 2024)	Pretrained PaliGemma2 (Steiner et al., 2024)	2.05	1.76	1.92	1.91
	FakeShield (Xu et al., 2025)	2.92	2.54	2.74	2.73
	MM-Det (Song et al., 2024)	3.18	2.79	2.96	2.98
	Grounded PaliGemma2 (no VOM)	3.47	3.21	3.34	3.34
	FORGE	**4.29**	**4.11**	**4.13**	**4.18**

```
{"region":"lighting","attribute":"consistent",
"polarity":"authentic"}]
Now extract claims from:
<explanation>
```

Results: Table 4 reports the three metrics across systems. FORGE cuts region hallucination by more than $2\times$ over the strongest baseline on both datasets and lifts forensic precision past 80%. Grounding PaliGemma2 on the forensic data alone (the ablation) halves the gap to FORGE but does not close it: RHR on DD-VQA falls from $\sim 40\%$ to 28.9% but only reaches 12.7% once the dense-patch VOM stream is added. The mask-based grounding baselines FakeShield and ForgerySleuth, despite producing spatial mask outputs, still hallucinate authentic regions as manipulated in 36-40% of cases on DD-VQA, text-level region grounding via the VOM stream more than triples their faithfulness in this respect. This is direct evidence that faithful region-grounding is not merely a matter of in-domain finetuning of an image-text encoder, nor of adding a spatial mask head; it requires the patch-resolution evidence stream the VOM provides.

Where the remaining hallucinations live: We manually inspect 50 FORGE errors per dataset to characterize residual failures. On DD-VQA, the dominant residual mode is over-extension within authentic faces: the model occasionally flags an authentic skin region as "slightly inconsistent" in addition to correctly identifying the true manipulated region (eyes or mouth). On DVF, residual hallucinations cluster around lighting and shadow claims on images with naturally complex lighting (e.g. outdoor scenes with low sun angles), where ground-truth annotators themselves required cross-reference to the source video to disambiguate. These patterns suggest that residual error is concentrated in scenes that are objectively ambiguous rather than in systematic failure modes of the model, and we treat them as targets for future work.

A3.4 Discussion

Three observations follow from the studies in this section.

Detection accuracy and explanation faithfulness can dissociate: BLIP-TI, FakeShield, and ForgerySleuth reach competitive detection accuracy (Table 1), yet Table 4 show that they hallucinate manipulated regions in 36-41% of explanations, with the ranking on RHR (FakeShield best, BLIP-TI worst) actively reversing the ranking on BLEU-4 (BLIP-TI best, ForgerySleuth worst). Detection-only and surface-only evaluations both produce orderings that explanation-quality evaluation rejects. A detection-only evaluation would treat these systems as near-peer with FORGE; the explanation evaluation does not.

Surface metrics preserve ordering but compress differences: The BLEU/ROUGE/CIDEr ranking in Table 1 agrees with our rubric judge on the top and bottom systems, but compresses the middle of the field. The BLIP-TI and grounded PaliGemma2 systems are essentially tied on CIDEr (within 0.1 points; PaliGemma2 marginally higher) and differ by ~ 3 BLEU-4 points and ~ 5 ROUGE-L points, yet differ by 0.55 on rubric helpfulness and 12 points on RHR. n-gram metrics see these systems as near-equivalent; explanation-quality metrics see them as substantially different. This reinforces NLG findings (Novikova et al., 2017) that n-gram metrics underweight content even when they preserve ordering.

The VOM stream is the source of the gain: The grounded PaliGemma2 ablation in Table A1 (FORGE minus the VOM stream) is the cleanest test of our central architectural claim. On every dimension of every study (human ratings, rubric ratings,

Table A6: **SOTA Comparison on Detection Performance**. The methods used for the comparisons are specifically designed for either face-manipulation or fully AI-generated content detection. However our generalized FORGE framework is designed for both face-manipulated and synthetic content detection. Furthermore, on the DF40 and DeMamba datasets, the competitor methods are tested in-domain, while for FORGE all evaluations performed are cross-domain. **Best** and second-best performances are marked.

Dataset	Method	Accuracy	AUC
Face-manipulated Data			
CelebDF (Li et al., 2020)	XceptionNet-BLIP-TI (Zhang et al., 2024)	62.41%	64.30%
	HifiNet (Guo et al., 2023)	67.20%	68.80%
	HifiNet-BLIP-TI (Zhang et al., 2024)	69.37%	71.00%
	RECCE (Cao et al., 2022)	67.96%	68.71%
	RECCE-BLIP (Zhang et al., 2024)	68.07%	68.36%
	RECCE-BLIP-TI (Zhang et al., 2024)	69.46%	70.21%
	TALL (Xu et al., 2023)	90.79%	
	ISTVT (Zhao et al., 2023)	84.10%	-
	Choi et al. (Choi et al., 2024)	89.00%	-
	$\mathcal{X}^2$-DFD (Chen et al., 2024b)	-	91.30%
	VLFFD (Sun et al., 2025a)	-	84.80%
	M2F2-Det (Guo et al., 2025)	-	95.10%
	FORGE	**92.86%**	**95.11%**
DF40 (Yan et al., 2024b)	$\mathcal{X}^2$-DFD (Chen et al., 2024b)	-	85.60%
	RECCE (Cao et al., 2022)	-	78.10%
	SBI (Shiohara and Yamasaki, 2022)	-	64.40%
	CORE (Ni et al., 2022)	-	76.10%
	IID (Huang et al., 2023)	-	75.70%
	UCF (Yan et al., 2023)	-	77.50%
	LSDA (Yan et al., 2024a)	-	77.80%
	CDFA (Lin et al., 2024)	-	75.90%
	ProgressiveDet (Cheng et al., 2024)	-	78.70%
	FORGE	**99.58%**	**98.96%**
Fully Synthetic Data			
DeMamba (Chen et al., 2024a)	TALL (Xu et al., 2023)	88.42%	-
	F3Net (Qian et al., 2020)	86.04%	-
	NPR (Tan et al., 2024)	83.45%	-
	STIL (Gu et al., 2021)	85.35%	-
	MINTIME-CLIP-B (Chen et al., 2024a)	89.98%	-
	FTCN-CLIP-B (Chen et al., 2024a)	89.67%	-
	CLIP-B-PT (Chen et al., 2024a)	41.82%	-
	DeMamba-CLIP-PT (Chen et al., 2024a)	79.98%	-
	XCLIP-B-PT (Chen et al., 2024a)	65.83%	-
	DeMamba-XCLIP-PT (Chen et al., 2024a)	79.31%	-
	XCLIP-B-FT (Chen et al., 2024a)	86.07%	-
	FORGE	**90.49%**	**92.67%**

RHR, AHR, FP) the VOM stream produces a separable gain over an MLLM finetuned on the same data with the same protocol; that gain also exceeds the gain delivered by mask-based grounding baselines (FakeShield, ForgerySleuth) over the same backbone family. The forensic explanation quality is attributable to the dense-patch evidence stream, not to scale, not to finetuning, not to data, and not to the choice of spatial output format.

A4 Detection Performance

On recent and popularly used datasets (both face-manipulated and fully synthetic), FORGE's DeepFake detection capabilities have been tested in Table A6. Since these datasets lack text annotations for the anomalies, the reasoning metrics for these datasets cannot be computed and thus only detection performance have been reported. Note that for DF40 (Yan et al., 2024b) and DeMamba (Chen et al., 2024a) datasets, all the competitor methods have been trained and tested on the same datasets (in-domain evaluation). However, FORGE has been tested in purely cross-domain setting, having been trained on DD-VQA (Zhang et al., 2024) and DVF (Song et al., 2024) still outperforms the state-of-the-art methods in detection performance.

The DeMamba results also provide a strong control for potential data-leakage concerns on our own DVF split. FORGE is trained on DVF and evaluated on DeMamba, an entirely different set of generators, and still outperforms competitors evaluated in-domain on DeMamba itself. Performance on a benchmark unseen during training cannot be explained by any within-dataset leakage, which further reinforces the generalization claims of Sec. **??**.

Table A7: **Ablation on data compression factors:** Results obtained by the FORGE model on different compression factors of the FF++ (Rossler et al., 2019) dataset using the same language annotations from DD-VQA (Zhang et al., 2024). The small differences across the compression factors highlight the robustness of FORGE.

Compression	Accuracy ↑	BLEU-3 ↑	BLUE_4 ↑	ROUGE-L ↑	CIDEr ↑
raw	94.17%	0.4587	0.4131	0.6270	2.6109
c23	93.61%	0.4610	0.4269	0.6103	2.6057
c40	94.12%	0.4649	0.4304	0.6285	2.6321

A5 Ablation on Resolutions

We evaluate the robustness of FORGE to varying levels of image compression using the FF++ dataset (Rossler et al., 2019) benchmark, considering the "raw", "c23", and "c40" compression settings commonly used in forensic literature. For this analysis, the language annotations (and data splits) are adopted from DD-VQA (Zhang et al., 2024) to provide a consistent evaluation protocol across all settings.

As summarized in Table A7, FORGE achieves stable performance across all compression levels, with only minimal fluctuations in detection accuracy and explanation metrics (BLEU-3, BLEU-4, ROUGE-L, CIDEr). Notably, the accuracy remains above 93.6% in all cases, and the scores on the language-based explainability metrics (e.g., BLEU-4 and CIDEr) remain consistently high, despite the potential visual degradation introduced by aggressive compression (e.g., "c40"). These results demonstrate that FORGE is highly robust to changes in input image quality, maintaining both detection effectiveness and the detailed quality of its region-grounded explanations, even under unfavorable, low-fidelity conditions.

This resilience is particularly important in practical forensic workflows, where manipulated content

is often subject to recompression or quality loss due to online sharing or platform-imposed constraints. By showing reliable performance across this spectrum, FORGE provides consistent and trustworthy DeepFake analysis regardless of compression artifacts or image degradation.

A6 More Qualitative Examples

To further illustrate the effectiveness of FORGE in generating precise, interpretable explanations, we present additional qualitative results on a variety of image types in Figure A1, including face-manipulated DeepFakes, fully AI-generated images, and authentic real-world images. These examples were selected to demonstrate both the breadth and subtlety of reasoning achieved by our method across manipulation scenarios.

Face-Manipulated Images: In face-manipulated examples, FORGE demonstrates its core strength: linking model predictions to region-specific visual evidence. For instance, when asked "Does the person's eyebrows look real or fake?", FORGE does not merely provide a binary answer, but instead grounds its explanation in visual attributes, remarking that the "eyebrows look a bit fake because they appear blurry, unclear or flawed," or "the eyebrows look fake, asymmetric, broken, and drawn on the face through computer editing." Similarly, when queried about the nose, the model observes that "the person's nose looks very fake, is unnaturally curved, and misaligned with the other facial features." When asked about the skin, the response cites "overly smooth skin lacking details." Such attribute-localized, textually explicit responses demonstrate not only detection accuracy, but also the interpretability and transparency required for practical DeepFake forensic workflows. By surfacing minute, manipulation-specific artifacts overlooked by conventional MLLMs, FORGE enables investigators to probe exactly where and how an image has been altered.

AI-Generated Images: For fully AI-generated content, FORGE fuses holistic scene analysis with attribute-level reasoning. In these examples, when asked whether the image appears real or fake, the model explicitly identifies telltale cues of synthesis: "the image looks fake," referencing "AI-generated appearance with unrealistic texture," or observing that "the person's hoodie looks mismatched with exaggerated colors." In another case, the model details that "the image looks fake; the person's eyes are asymmetric, the nose looks curved, and the mouth is misaligned with other facial features," clearly indicating knowledge of how AI-generation often fails to capture coherent facial structure. These multi-faceted, region-specific explanations set FORGE apart from prior methods that offer only undifferentiated global judgements, and highlight the model's utility for both forensic screening and downstream content auditing.

Real Images: For authentic images, FORGE maintains high precision and does not falsely attribute benign artifacts as fake. Its answers highlight the absence of manipulation with clarifying details: acknowledging "naturally even skin, bushy eyebrows, and a full mouth," or commenting that "the person has proper hair and hairstyles," "illuminated skin, sparkling eyes, and a broad nose." This restraint, coupled with accurate grounding, ensures trustworthiness and minimizes false alarms in real-world deployment scenarios.

Analysis and Implications: These qualitative results underscore several key strengths of FORGE. First, the model's ability to align its explanations with concrete, localized regions in the image ensures that each rationale is visually and forensically grounded. This explicit region-level correspondence builds trust in the model's predictions and directly addresses the limitations of prior methods that rely solely on global analysis or binary outputs.

Second, FORGE demonstrates remarkable flexibility in handling both open-ended and attribute-specific forensic queries. Whether prompted about the overall authenticity of an image ("Does the image look real or fake?") or about specific facial features (such as `eyebrows`, `nose`, or `skin`), FORGE provides granular, context-aware explanations. For face-manipulated images, the system surfaces manipulation artifacts like blurry or asymmetric eyebrows, unnaturally curved noses, or unnaturally smooth skin, clearly indicating both the presence and nature of DeepFake manipulations. For AI-generated content, FORGE is able to synthesize holistic scene cues with localized anomalies, referencing unnatural textures, mismatched colors, and misaligned facial features. Real images are correctly described with natural, harmonious feature attributions, showing the system's precision and its resistance to false positives.

A particularly notable advantage, as evidenced across these examples, is FORGE's competency in addressing fine-grained, query-driven prompts. By

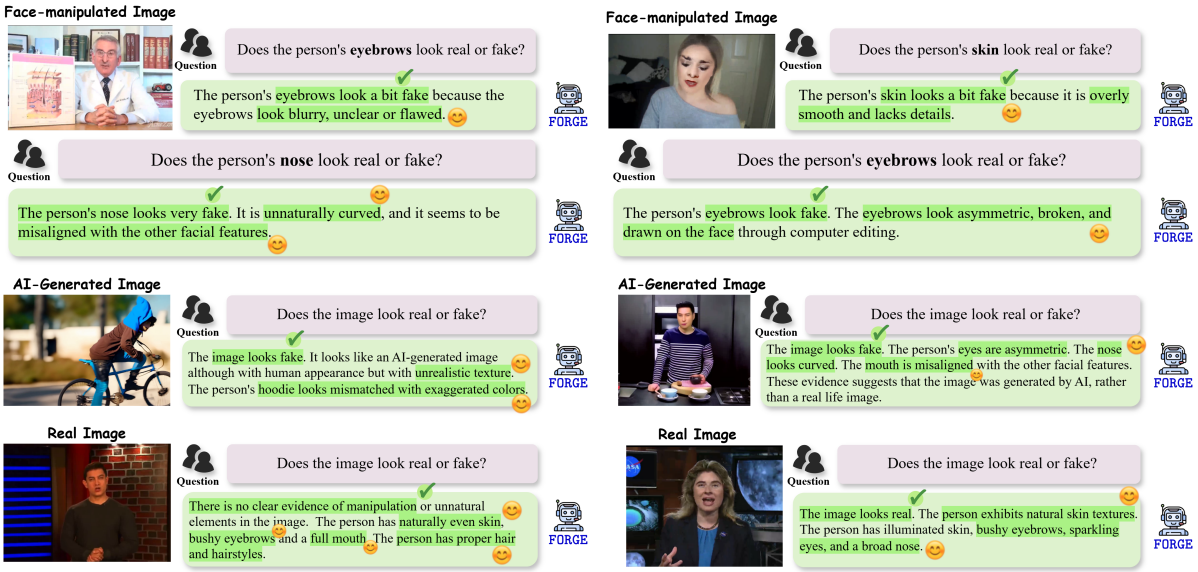


Figure A1: **Qualitative examples** demonstrating FORGE's detailed, region-grounded explanations across diverse image types. For face-manipulated images, FORGE identifies subtle inconsistencies such as blurry or flawed eyebrows, unnatural nose shapes, and misaligned facial features. In AI-generated images, it captures both global and local anomalies, including unrealistic textures, mismatched colors, and asymmetric facial components. For authentic images, the model confirms natural, harmonious features without producing false positives. Each example illustrates FORGE's capability to answer fine-grained, attribute-specific queries, providing forensic insights that enhance transparency and interpretability well beyond traditional binary classification methods.

Table A8: Evaluation of FORGE stratified by finegrained region or scene aspect.

Dataset	Region/Aspect	Accuracy	AUC	BLEU-4	ROUGE-L	CIDEr
DD-VQA (Zhang et al., 2024)	Eyes	95.82	96.46	0.4521	0.6452	2.7813
	Nose	95.14	96.11	0.4463	0.6398	2.6954
	Mouth	94.61	95.63	0.4387	0.6331	2.6612
	Eyebrows	92.37	94.17	0.4015	0.6024	2.4217
	Skin	93.75	95.45	0.4249	0.6196	2.5839
	Overall	94.12	95.39	0.4304	0.6285	2.6321
DVF (Song et al., 2024)	Background	95.35	96.67	0.4412	0.5487	2.2615
	Lighting	94.94	96.19	0.4375	0.5421	2.2328
	Shadows	94.37	95.71	0.4318	0.5364	2.2012
	Textures	93.89	95.23	0.4236	0.5279	2.1527
	Reflections	93.10	94.74	0.4182	0.5213	2.1049
	Object Boundaries	94.12	95.43	0.4291	0.5339	2.1853
	Color Consistency	93.57	95.01	0.4210	0.5260	2.1396
	Overall	94.47	95.82	0.4279	0.5338	2.1744

moving beyond simple authenticity judgments, and providing attribute-resolved rationales, the model empowers forensic experts with actionable insights, identifying precisely which regions or features are manipulated and explaining why.

Collectively, these qualitative results highlight FORGE's unique contributions: robust region-level grounding, the ability to deliver interpretable and visually anchored rationales, and consistent performance across both fine-grained and holistic queries. By bridging the gap between detection accuracy and explanation transparency, FORGE advances the state of the art and offers a practical, reliable tool for explainable DeepFake forensics.

A7 Stratified Evaluation by Region and Aspect

The region-referential evaluation reports aggregate faithfulness and hallucination metrics. In Table A8 we report FORGE's performance broken down by the specific facial region (DD-VQA) or scene aspect (DVF) that each query targets, providing a finer-grained view of where the model is strong and where residual error concentrates. Table A8 reports accuracy, AUC, and the three generation metrics per region and aspect.

On DD-VQA, performance is highest on the regions that carry the most discriminative manipulation evidence: eyes (95.82% accuracy) and nose (95.14%), both of which exhibit pronounced geometric and blending artifacts under face manipulation. Performance is lowest on eyebrows (92.37%), consistent with their smaller spatial extent and the subtlety of eyebrow-level edits. On DVF, the model is strongest on background and lighting (95.35% and 94.94%), where synthesis artifacts are spatially broad, and weakest on reflections (93.10%) and color consistency (93.57%), which require global cross-region comparison and are the aspects most often ambiguous even to human annotators. Across both datasets the spread between the best and worst region is under 3.5 accuracy points, indicating that the dense-patch stream surfaces evidence consistently rather than concentrating performance on a single easy region.