FAKECUES: LEARNING SCENE-CONDITIONED CUES FOR AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Advances in generative models enable high-fidelity image synthesis across diverse scenes, characterized by visual style and subject content, challenging reliable AI-generated image detection. Recent multimodal detectors identify visual evidence for their decisions but do not explicitly model its scene-dependent diagnostic value: textures or structures that suggest AI generation in photographs may reflect ordinary artistic expression in paintings, potentially leading to misclassification. To address this limitation, we introduce FakeCueS, a framework that uses scene context to guide forensic cue generation and evaluation. Predicted scene representations guide the model to generate context-specific cues. An online cue evaluator verifies visual support and groups semantically related cues to assess their source association within each scene. These assessments provide set-level rewards for group-relative policy learning alongside image-level supervision, without reference rationales or region annotations. We further introduce RedScenes, a dataset of 168,688 social-media images spanning 70 scenes. FakeCueS achieves 91.53% accuracy on RedScenes, exceeding the strongest evaluated zero-shot baseline by 12.25%. Without target-domain fine-tuning, it achieves 94.49% and 83.09% accuracy on AnomReason-Deepfake and Chameleon, respectively.

1 INTRODUCTION

Generative AI enables high-quality image creation (Ho et al., 2020; Rombach et al., 2022) while allowing fabricated events to circulate as authentic visual records. Reliable AI-generated image detection supports content moderation and information verification (Zhou et al., 2025). Beyond a label or confidence score, users need visual evidence to examine decisions (Ji et al., 2026b). Multimodal large language models (MLLMs) support such explainable detection by combining source prediction with natural-language explanations (Zhou et al., 2025; Kang et al., 2025).

However, image-level source labels alone do not ensure that their explanations describe visible image content or identify cues that distinguish real from AI-generated images. Recent methods address these limitations through explanation supervision and preference optimization (Zhou et al., 2025), region-level grounding (Ji et al., 2026b), and verification of visual accuracy and synthetic specificity (Deng et al., 2026). These approaches extend supervision beyond source classification to the forensic cues used to justify predictions.

Despite these advances, cues learned and validated on certain types of images may not retain their diagnostic value on others. For example, some forensic training datasets focus on photographic images and filter out non-photorealistic content, including anime, cartoons, or digital art (Kang et al., 2025; Deng et al., 2026), leaving a coverage gap for detection across visual styles. Figure 1 illustrates how applying photographic criteria to stylized content can lead models to mistake ordinary visual expression for generation artifacts. Scene context is thus crucial for assessing whether an observation provides reliable evidence, including not only visual style, which shapes textures and forms, but also subject content, which guides structural and semantic interpretation. Accounting for this dependence is essential to learning and evaluating valuable forensic cues.

To address this challenge, we first construct RedScenes, providing source, visual-style, and subject annotations for 151,790 training and 16,898 test images. It spans photography, anime, 3D rendering, graphic design, and painting, with multi-labeling human, animal, text, and other subject. Every style includes Real and Fake images.

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



Figure 1: **Scene-dependent cue validity.** Cues that suggest AI generation in photographs may reflect ordinary stylization in art. Both baselines misclassify the Real examples in (b) and (c), while FAKECUES correctly classifies all three. For clarity, Fake denotes images generated or edited with AI (AIGC), and Real denotes non-AI-generated images (Non-AIGC), including camera images and human-created artwork.

Building on RedScenes, we propose FAKECUES to learn open-ended forensic cues through scene-guided generation and evaluation. A predicted scene representation encoding visual style and subject content conditions both cue generation and source classification from the input prefix. A frozen vision-language verifier checks the visual support of observations, while a scene-conditioned cue scoring groups similar cues online and assesses their association with source. These assessment, together with semantic novelty, define set-level rewards for group-relative policy learning (Shao et al., 2024), alongside image-level source and scene supervision, without per-image rationale annotations.

Overall, our main contributions are:

- We propose FAKECUES, a scene-guided framework for AI-generated image detection that incorporates visual style and subject context into forensic cue generation and evaluation. It learns open-ended cues via visual verification and scene-conditioned source association, without per-image rationales.
- We build RedScenes, a dataset of 168,688 social-media images with joint source, style, and multilabel subject annotations. Its five visual styles and 14 subject-label combinations form 70 scenes, supporting AI-generated image detection across styles and subjects.
- Experiments show that FAKECUES leads the evaluated baselines with 91.53% on RedScenes. Cross-dataset evaluation further demonstrates the generalization of our scene-conditioned cue learning framework, achieving 94.49% on AnomReason-Deepfake and 83.09% on Chameleon.

2 RELATED WORK

2.1 AI-GENERATED IMAGE DETECTION

CNN-based detectors study cross-generator transfer (Wang et al., 2020). Detection uses diffusion or autoencoder reconstruction errors (Wang et al., 2023; Ricker et al., 2024), texture contrasts (Zhong et al., 2024), and local pixel relations (Tan et al., 2024). Pretrained CLIP (Radford et al., 2021) enables frozen-feature classification (Ojha et al., 2023) and forgery-aware adaptation (Liu et al., 2024). TAP learns attention pooling over frozen patch tokens (Abdullah et al., 2026); SimLBR regularizes the decision boundary around real images through latent blending (Dhakal et al., 2026). DRCT uses diffusion reconstructions for contrastive training (Chen et al., 2024), while PPL distributes learning across patches (Yang et al., 2026). HiDA-Net combines native-resolution tiles and global views (Mu et al., 2026). MCAN's encoder adapter aggregates spatial, frequency, and chromaticity cues (Tan et al., 2026c).

Multimodal detectors explain source judgments (Wen et al., 2025). Forensic encoders and adapters fuse low-level and semantic features (Zhou et al., 2025; Wang et al., 2026; Sun et al., 2025). Forgery prompt learning maps detector outputs to textual-response prompts (Yu et al., 2025). Region-grounded models predict labels, locations, and explanations (Huang et al., 2025; Kang et al., 2025).

Contextual guidance uses manipulation-domain tags (Xu et al., 2025), scene-specific experts (Hu et al., 2026), and object-category structural checks (Ji et al., 2025). ReAlign aligns visual features with reasoning text but classifies without generating explanations at inference (Huang et al., 2026). Fake-HR1 learns to choose direct answers or explicit reasoning (Jiang et al., 2026b).

Supervision includes attribute-based face-forgery VQA (Zhang et al., 2024) and reference explanations with preference optimization (Zhou et al., 2025). Descriptor selection uses real/fake occurrence differences (He et al., 2025) or feature-question accuracy (Chen et al., 2025b). Reinforcement learning rewards answers and format (Li et al., 2025), region alignment (Ji et al., 2026b), and reference-based re-examination (Ji et al., 2026a). Veritas adds face-forgery planning and reflection feedback (Tan et al., 2026b). FakeVLM-R1 uses common-sense counterevidence and logical-consistency feedback (Zhu et al., 2026). DeepfakeJudge scores complete explanations given image and source label (Kuckreja et al., 2026), whereas Defake-o3 trains localized verification of visual accuracy and synthetic specificity (Deng et al., 2026).

2.2 Datasets for AI-Generated Image Detection

GenImage tests cross-generator and degraded-image detection (Zhu et al., 2023). Community Forensics spans thousands of generators (Park & Owens, 2025). Chameleon targets generated images humans mistake for real (Yan et al., 2025a). SID-Set and DFBench cover real, fully generated, and locally manipulated images, plus degraded real images in DFBench (Huang et al., 2025; Wang et al., 2025a). Coverage includes remote sensing and documents (Wen et al., 2025), media modalities, and forensic tasks (Ye et al., 2025; Wang et al., 2025b).

Annotations include source labels and generated explanations (Huang et al., 2025; Wen et al., 2025), SynthScars' irregular artifact masks (Kang et al., 2025), FakeXplained's boxes and explanations (Ji et al., 2026b), and X-AIGD's hierarchical artifact categories with pixel-level annotations (Xiao et al., 2026). AnomReason evaluates semantic anomaly detection and reasoning (Tan et al., 2026a). HEIE studies visual implausibility with heatmaps, explanations, and plausibility scores (Yang et al., 2025). RedScenes jointly labels source, style, and subject in photographic and non-photographic images.

3 RedScenes Dataset

We introduce RedScenes, a dataset of 168,688 social-media images with source, style, and subject annotations for scene-aware AIGC detection (Figure 2).

3.1 Image Collection and Source Annotation

Each sampled public RedNote (Xiaohongshu) post contributes its first image, unless a conspicuous text watermark is detected. Platform AIGC tags define candidate pools; untagged posts are randomly sampled, with additional portrait-photography images. Perceptual hashing and local matching remove duplicates. Source labels undergo human review followed by a quality check; only agreed cases are retained. Annotators can consult post content, author history, creation statements, and platform tags. The Fake source includes full generation, local AI edits, and AI-assisted drawings; the Real source includes camera images and human-created artwork. Appendix A details the collection and annotation protocol.

3.2 Scene Definition and Annotation

Each image has one style (photography, anime, 3D rendering, graphic design, or painting) and one or more subject labels (human, animal, other, or prominent text). A style paired with the complete subject-label set defines a scene combination. Individual style–subject pairs form overlapping scene components for cue verification. Gemini 3 Pro generates these labels from the image, its confirmed source label, and category definitions. These annotations provide scene supervision without manual review or correction.

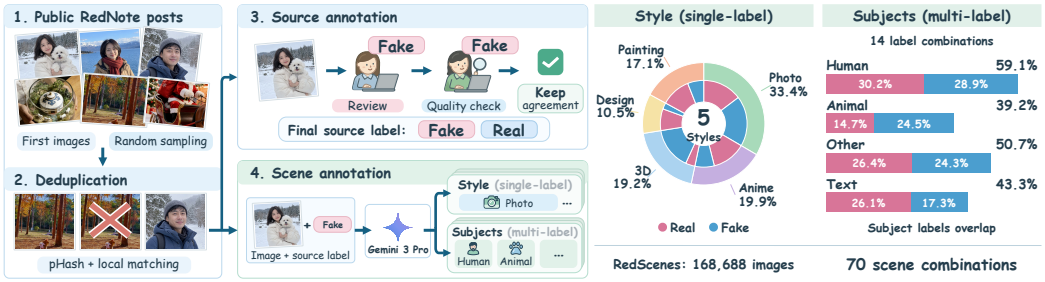

Figure 2: **RedScenes construction and composition.** Left: collection and annotation. Right: composition of all 168,688 images. The outer ring shows visual styles, and the inner ring divides each style into Real and Fake. Subject bars show overlapping label frequencies, split by source. All percentages use the full dataset as the denominator. Portrait examples are AI-generated illustrations.

3.3 Composition and Splits

The splits preserve the original image-level assignments. Within each split, Real photography and graphic-design images are downsampled within strata defined by subject-label combinations. The resulting training split contains 151,790 images, with 75,895 per source class; the test split contains 16,898 images, with 8,449 per class. Both splits cover five styles, 14 subject-label combinations, 70 scene combinations, and all 20 scene components. Both source classes occur in every style, with varying proportions (Figure 2). Sampling details and counts appear in Appendix A.3.

4 Method

FakeCues combines scene-conditioned source classification with open-ended visual cue learning (Figure 3). Its shared backbone and separate frozen visual verifier use Qwen3.5-4B (Qwen Team, 2026). Given image x and instruction q, the model predicts source and generates cues linking visual observations to source judgments. Image-level labels supervise source, style, and subject prediction. A set-level reward trains cue generation without reference annotations. We encode $y = 1$ for Fake and $y = 0$ for Real.

4.1 Scene Conditioning

The diagnostic value of a visual cue can vary with image style and subject content. Scene conditioning supplies this context to source classification and cue generation. Let $\mathbf{v}$ be the mean of the image encoder's output features. The functions f_{style} and f_{subject} predict style and subjects from $\mathbf{v}$. A scene mapping combines these predictions with visual features:

$$\mathbf{s} = f_{\text{scene}}([\mathbf{v}; f_{\text{style}}(\mathbf{v}); f_{\text{subject}}(\mathbf{v})]) , \tag{1}$$

where semicolons denote concatenation and f_{scene} maps to the token embedding dimension.

For embedding $\mathbf{e}_t$ at position t in the image-and-instruction prefix, the scene vector adds a bias scaled by fixed coefficient α. Mean pooling the backbone's final hidden states over this conditioned prefix yields global representation $\mathbf{h}$. The source classifier f_{src} takes $[\mathbf{h}; \mathbf{s}]$ and predicts Fake probability p_{src}:

$$\widetilde{\mathbf{e}}_t = \mathbf{e}_t + \alpha\mathbf{s}, \qquad p_{\text{src}} = f_{\text{src}}([\mathbf{h}; \mathbf{s}]). \tag{2}$$

Each backbone-generated cue contains an *observation* describing a visible fact, *reasoning* explaining its source implications, and a *verdict* stating the supported class. The concrete observation enables image-specific verification, while a *summary* of the visual phenomenon supports semantic grouping and historical statistics across images.

4.2 Cue Scoring

The cue reward combines visual support, source association, and semantic novelty. The frozen BGE-base-en-v1.5 encoder (Xiao et al., 2024) embeds cue summaries; we group them into semantic

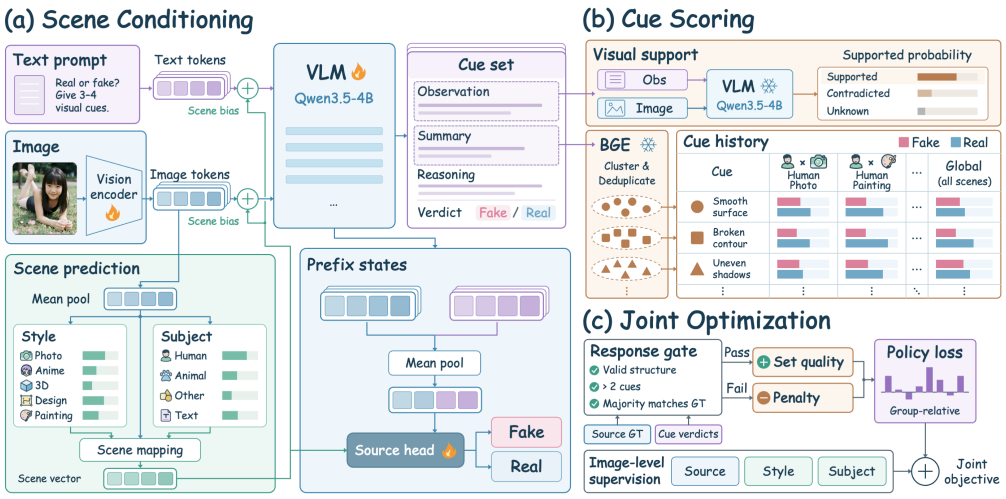

Figure 3: **Overview of FAKECUES.** (a) Predicted style and subject context guides cue generation and source prediction. (b) Visual support and scene-conditioned statistics score the cues. (c) Cue-set rewards and image-level supervision jointly train the model.

clusters to limit repeated evidence. Here $k(c)$ denotes cue c's cluster. Summaries in response cue set C are processed in generation order, compared with retained representatives, then matched to historical clusters or assigned new ones. Keeping the first cue per final cluster yields representative set $\widetilde{C}$.

Visual support. The independent verifier receives the image and cue observation, without the cue's source verdict. It scores supported, contradicted, and unknown options. The softmax-normalized probability of supported, $g(x, c)$, provides a continuous support weight.

Source association. Historical reporting rates associate visual phenomena with source within each scene s, defined by a training style–subject label pair. Counts for cluster k decay over image visits from its discovery. Let $\rho_k^{(y)}(s)$ denote the weighted fraction of source-y visits in scene s reporting cluster k. Reports across candidates for one image count once. The source-class contrast is

$$\Delta_k(s) = \frac{\rho_k^{(1)}(s) - \rho_k^{(0)}(s)}{\rho_k^{(1)}(s) + \rho_k^{(0)}(s)}. \tag{3}$$

Positive values indicate more frequent reports for Fake images, while negative values indicate the opposite. Scene scores blend with the global score according to statistical support, relying more on global history when support is low. Averaging the blended scores across the image's subject-associated scenes yields association score $a_k(x)$. Clusters without usable history receive zero association scores. Appendix B.3 defines counts and backoff weights.

Set reward. A novelty bonus encourages exploration of less frequently reported phenomena. It decreases with N_k, the number of distinct historical images reporting cluster k. Cue verdict $y_c \in \{0, 1\}$ uses the source encoding of y, and K is the minimum count for normalizing set quality:

$$Q(C, x) = \frac{1}{\max(K, |\widetilde{C}|)} \sum_{c \in \widetilde{C}} g(x, c) \left[(2y_c - 1) a_{k(c)}(x) + \frac{\lambda_{\text{nov}}}{\sqrt{1 + N_{k(c)}}} \right]. \tag{4}$$

The signed association term is positive when the verdict agrees with the historical source tendency, and λ_{nov} controls the exploration bonus. The denominator is bounded below by K: responses with fewer than K distinct representatives still divide by K, reducing the incentive to return only a few high-scoring cues. Semantic duplicates receive no additional credit. All candidates share the history before the image visit, updated only after the entire group is scored.

Table 1: In-domain accuracy (%) on RedScenes (16,898 images). Real/Fake columns report class recall. Bold/underline indicate the best/second-best result in each column.

Method	Overall	Class		Style				
		Real	Fake	Photo	Anime	3D	Design	Painting
Classification-based								
UnivFD (ViT-L/14)	46.01	74.49	17.52	41.86	56.25	26.72	68.61	50.03
NPR (ProGAN-4class)	50.83	**95.55**	6.11	46.11	61.67	22.06	77.17	63.66
PGC (SDv1.4)	51.73	76.28	27.17	49.65	55.63	37.35	69.35	56.64
AIDE (ConvNeXt-XXL)	55.18	74.23	36.12	53.71	58.90	44.98	66.91	57.99
Effort (SDv1.4)	57.72	55.18	60.27	55.23	58.73	62.84	61.19	53.58
B-Free (DINOv2-B/14-reg4)	59.36	40.61	78.12	65.09	46.07	71.52	55.64	52.22
DDA (DINOv2-L/14)	62.53	48.41	76.65	65.88	57.18	71.21	62.83	52.22
MLLM-based								
FakeShield (LLaVA-v1.5-13B)	45.12	71.06	19.17	48.50	43.60	29.41	57.28	50.45
FakeReasoning (LLaVA-13B)	51.89	39.59	64.19	50.60	41.99	66.91	53.37	48.16
FakeVLM (LLaVA-v1.5-7B)	57.80	19.60	96.00	61.93	43.18	80.72	46.80	47.71
UniGenDet (BAGEL-7B-MoT)	58.62	18.38	<u>98.86</u>	71.43	38.42	81.77	42.66	40.76
AIGI-Holmes (LLaVA-1.6-Mistral-7B)	59.77	74.45	45.09	58.67	62.21	52.73	70.82	60.22
General-purpose MLLMs								
InternVL3.5-8B (Zero-shot)	60.59	46.11	75.06	68.48	40.08	75.22	63.97	50.45
GLM-4.6V-Flash (Zero-shot)	65.71	79.12	52.29	59.86	60.36	67.16	<u>81.59</u>	<u>72.06</u>
Qwen3.5-4B (Zero-shot)	66.92	34.57	**99.27**	87.53	43.87	82.73	50.48	45.59
Gemini 3 Flash (Zero-shot)	71.55	45.71	97.40	81.77	63.58	85.76	62.04	50.63
Qwen3.7-Plus (Zero-shot)	77.38	56.67	98.08	<u>90.28</u>	68.76	<u>89.13</u>	66.86	55.28
Qwen3.8-Max (Zero-shot)	<u>79.28</u>	64.97	93.60	89.93	<u>71.11</u>	87.92	76.43	59.90
FakeCueS (Qwen3.5-4B)	**91.53**	<u>90.67</u>	92.39	**93.89**	**90.14**	**91.01**	**90.08**	**89.99**

4.3 Joint Optimization

For each training image, the policy samples G responses o_i containing cue sets C_i. A response receives $R_i = Q(C_i, x)$ when its structure is valid, its cue count before deduplication is within the prescribed range, and a strict majority of deduplicated verdicts agree with image label y. Otherwise, the failed condition determines a negative reward (Appendix B.4).

We use group-relative advantages (Shao et al., 2024) in a REINFORCE-style policy objective (Williams, 1992), normalizing rewards within each image's response group:

$$A_i = \frac{R_i - \bar{R}}{s_R + \epsilon}, \qquad \mathcal{L}_{\text{pg}} = -\frac{1}{G}\sum_{i=1}^{G} A_i \log \pi_\theta(o_i \mid x, q). \tag{5}$$

Here $\bar{R}$ and s_R are the within-image reward mean and sample standard deviation, and ϵ ensures numerical stability. Policy π_θ with parameters θ uses the same sampling and likelihood distribution, with log probability covering the complete response. Advantages remain constant during optimization and become zero when reward variation is negligible.

Image-level losses comprise source binary cross-entropy $\mathcal{L}_{\text{source}}$, style cross-entropy $\mathcal{L}_{\text{style}}$, and subject multilabel binary cross-entropy $\mathcal{L}_{\text{subject}}$. The joint objective is

$$\mathcal{L} = \mathcal{L}_{\text{pg}} + \mathcal{L}_{\text{source}} + \lambda_{\text{scene}}(\mathcal{L}_{\text{style}} + \mathcal{L}_{\text{subject}}), \tag{6}$$

where λ_{scene} weights scene supervision. Both objectives update the backbone's LoRA adapters (Hu et al., 2022), with pretrained weights frozen. Image-level supervision also updates the source head and scene modules. Inference requires one prefix forward pass for source classification. We train with AdamW (Loshchilov & Hutter, 2019) using DeepSpeed (Rasley et al., 2020) (Appendix B).

5 EXPERIMENTS

We adapt Qwen3.5-4B on 151,790 RedScenes training images for one epoch. Unless specified, detection uses *Direct* prediction: the source head outputs a probability from the input prefix without autoregressive generation (details in Appendices B and D).

5.1 IN-DOMAIN EVALUATION

Table 1 compares detectors and MLLMs on 16,898 test images balanced by source. FakeCueS reaches 91.53% accuracy, leading all five styles. Anime and Painting reach 90.14% and 89.99%, exceeding their respective strongest baselines by 19.03 and 17.93 percentage points. Both require distinguishing human artwork from generated imagery despite their non-photographic appearance. The 3.90-point range from Painting (89.99%) to Photo (93.89%) indicates consistent performance across styles.

Adaptation primarily corrects false alarms: untuned Qwen3.5-4B recognizes 99.27% of Fake images but only 34.57% of Real images. FakeCueS raises Real recall to 90.67%, while Fake recall becomes 92.39%. This tradeoff benefits RedScenes, whose Real class includes both artwork and photographs. Appendix D.2 gives style–subject breakdowns.

5.2 OUT-OF-DOMAIN EVALUATION

Table 2: Out-of-domain accuracy (%). Prior results are quoted from the cited papers. Bold and underlined scores denote the best and second-best listed results.

(a) AnomReason-Deepfake

Method	Acc. ↑
General-purpose MLLMs	
LLaVA-OV-7B	66.26
InternVL3-8B	62.84
Qwen2.5-VL-7B	65.41
Qwen2.5-VL-72B	77.60
Gemini-2.5-pro	85.65
GPT-o3	85.60
GPT-4o	87.76
Specialized detectors	
AnomReasonor-7B	82.61
UniGenDet†	88.59
SPARED Iter3	<u>92.18</u>
FakeCueS	**94.49**

(b) Chameleon

Method	Acc. ↑
Classifier-based detectors	
CNNSpot	60.89
UnivFD	60.42
DIRE	57.83
PatchCraft	55.70
NPR	57.81
AIDE	65.77
SynerDetect	79.42
MLLM-based detectors	
FakeShield	58.70
Ivy-xDetector	73.17
TranX-Adapter	**85.10**
FakeCueS	<u>83.09</u>

General-purpose MLLMs and AnomReasonor follow the AnomReason benchmark; † is reported by SPARED. Chameleon classifier baselines use AllGenImage training. Sources, training settings, and released-set details are given in Appendix D.1.

We evaluate the same checkpoint without target-domain fine-tuning on released AnomReason-Deepfake (Tan et al., 2026a) (22,752 images) and Chameleon (Yan et al., 2025a) (26,033 images) test sets. Table 2 compares published references under their respective training settings (Appendix D.1). FakeCueS reaches 94.49% accuracy on AnomReason-Deepfake, above the listed reference scores. Real and Fake recall are 98.50% and 90.03%, respectively. Most remaining errors are missed generated images.

Chameleon's generated images were selected to challenge human perception and pose harder transfer for FakeCueS. Accuracy is 83.09%, above SynerDetect's 79.42% but below TranX-Adapter's 85.10%. Real and Fake recall are 86.83% and 78.10%, respectively. Both classes are harder to recognize than on AnomReason-Deepfake; missed Fake images again dominate the errors, identifying generated content as the main transfer challenge.

5.3 Ablation Studies

Table 3 compares complete training configurations, including prompts and prediction paths. *Classifier with Scene* retains scene conditioning and image-level supervision without cue generation; the two FakeCueS rows share one checkpoint.

Table 3: Accuracy (%) by training and prediction setup.

		RedScenes (ID)			AnomReason-Deepfake (OOD)		
Training setup	Prediction	Overall	Real	Fake	Overall	Real	Fake
Zero-shot	Label decoding	66.92	34.57	**99.27**	92.37	93.48	<u>91.12</u>
+ Cue GRPO	Cue voting	89.55	85.19	<u>93.90</u>	93.50	97.92	88.59
+ Classifier with Scene	Source head	**91.73**	**94.39**	89.08	92.26	<u>98.49</u>	85.34
FakeCueS (joint)	Source head	<u>91.53</u>	<u>90.67</u>	92.39	<u>94.49</u>	**98.50**	90.03
	Cue voting	90.36	88.03	92.69	**94.85**	96.88	**92.60**

"+" denotes adaptation of Qwen3.5-4B. FakeCueS shares one checkpoint across Direct (source head) and Cue-based (cue voting) prediction. Real/Fake accuracy is class recall. Bold/underline: best/second-best full-set results. Settings: Appendix D.1.

Joint training for detection. *Classifier with Scene* and FakeCueS Direct reach 91.73% and 91.53% on RedScenes. On AnomReason-Deepfake, Direct reaches 94.49% versus the classifier's 92.26%. Fake recall increases from 85.34% to 90.03%, while Real recall stays nearly unchanged at 98.49% and 98.50%. Despite similar ID accuracy, the joint model recognizes more generated images in this external test without increasing false alarms.

Cue-based prediction. *Cue-based* accuracy on RedScenes is 90.36%, above *Cue GRPO*'s 89.55% but below Direct's 91.53%. On AnomReason-Deepfake, voting exceeds Direct (94.85% vs. 94.49%): Fake recall rises from 90.03% to 92.60%, while Real recall falls from 98.50% to 96.88%.

5.4 Cue Analysis

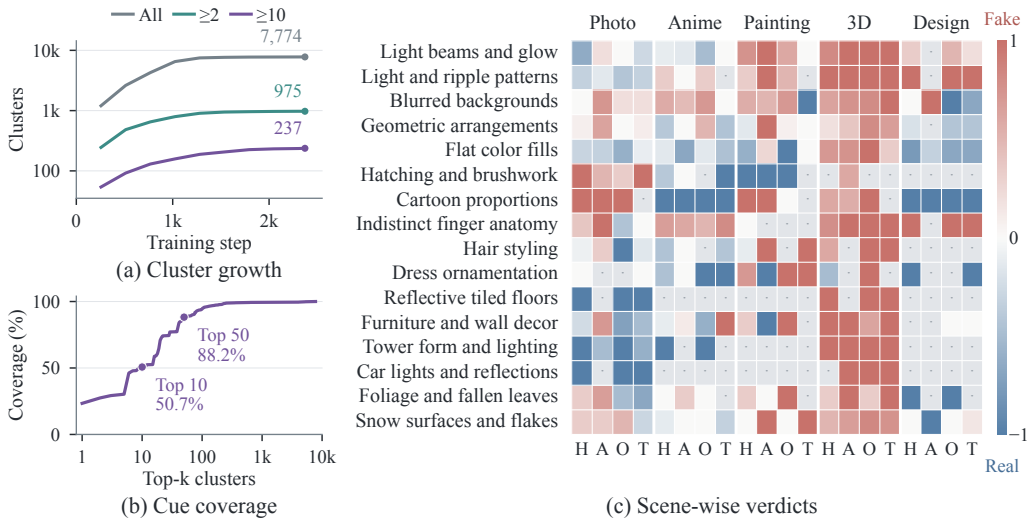

Figure 4: **Cue clusters during training.** (a) Growth and reuse. (b) Image coverage by the most-reused clusters. (c) Mean cue verdicts across scenes: red favors Fake, blue favors Real; grey denotes no retained examples. H/A/O/T denote human, animal, other, and text. All panels use rank-0 history.

Cluster growth and reuse. Figure 4(a,b) tracks rank-0 cluster growth and reuse. At step 2,372, the 7,774 clusters include 975 seen on at least two distinct images and 237 on at least ten. The top 10 and 50 clusters cover 50.7% and 88.2% of recorded images, respectively, counting each image once. Broad reuse thus concentrates in a small vocabulary amid a long tail of infrequent clusters. The remaining 87.5% of clusters appear in only one image in this history.

Scene-dependent verdicts. Figure 4(c) summarizes model verdicts over selected retained examples. *Hatching and brushwork* receives Real verdicts in Painting–Animal but predominantly Fake verdicts in 3D–Animal. Within Photo, *furniture and wall decor* shifts from predominantly Real for Other to Fake for Animal. Shared descriptions thus carry different source judgments across these scene contexts. The variation spans both visual styles and subject categories within a style. Appendix D.3 details selection and aggregation.

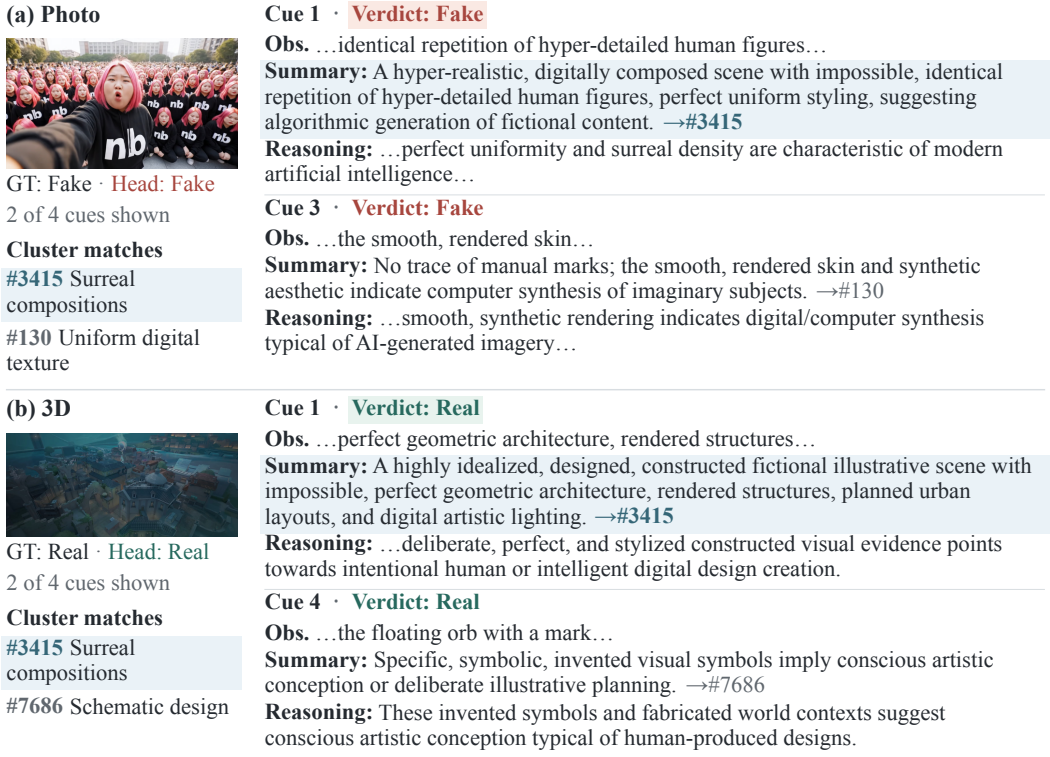

Figure 5: **One matched cluster, opposite verdicts.** Two of four cues are shown per image. Blue highlights mark complete summaries matched post hoc to frozen cluster #3415; ellipses mark omitted observation or reasoning text.

Qualitative examples. In Figure 5, repeated people in Photo and idealized architecture in 3D match the same *surreal compositions* cluster post hoc, yet accompany opposite model verdicts. Cues refer to rendered skin and a floating symbol: the reusable description groups related phenomena, while image-specific reasoning interprets them within the scene. Both source-head predictions match the ground truth in these selected examples.

6 Conclusion

In this paper, we present FakeCueS, a scene-guided framework for AIGC detection. It combines scene-conditioned cue generation with online assessment for cue learning. Moreover, we introduce RedScenes to support training across diverse scenes. Experiments demonstrate strong performance and cross-dataset generalization, showing the role of scene context in forensic cue interpretation.

AI USE STATEMENT

Gemini 3 Pro was used to generate visual-style and subject annotations; Real/Fake source labels were assigned by human annotators. Codex assisted with literature organization, method and experiment discussion, implementation inspection and revision, English drafting and translation, LaTeX assembly, and figure development. Local checks included comparisons between equations and code, citation checks, and PDF compilation and visual inspection. The authors reviewed and verified all AI-assisted content and take responsibility for the final manuscript.

ETHICS STATEMENT

This study uses images from public Xiaohongshu posts solely for research on AI-generated image detection, without user identification or profiling. The retained dataset consists only of images and labels. Any public release will require prior legal review and clearance regarding data-use permissions, copyright, and privacy.

REPRODUCIBILITY STATEMENT

The construction of RedScenes, model objectives, and evaluation protocols are specified in Sections 3–5. Implementation settings and statistical estimators are detailed in Appendices B and B.3.

REFERENCES

Ahmed Abdullah, Nikolas Ebert, and Oliver Wasenmüller. TAP into the patch tokens: Leveraging vision foundation model features for AI-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 379–389, 2026.

Yicheng Bao, Xiahui Guo, Xuhong Wang, and Xin Tan. SPARED: Reasoning-based AI-generated image detection via adversarially edited data. *arXiv preprint arXiv:2608.12876*, 2026.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7621–7639, 2024.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes AI-generated image detector easier generalizable. In *Advances in Neural Information Processing Systems*, volume 38, pp. 106475–106500, 2025a.

Yize Chen, Zhiyuan Yan, Guangliang Cheng, Kangran Zhao, Siwei Lyu, and Baoyuan Wu. X2-DFD: A framework for explainable and extendable deepfake detection. In *Advances in Neural Information Processing Systems*, volume 38, pp. 83792–83839, 2025b.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, pp. 2632–2652, 2024.

Bowen Deng, Jiahui Zhan, Yikun Ji, Haozhen Yan, and Jianfu Zhang. Defake-o3: From speculative rationales to verifiable evidence for explainable AIGI detection. *arXiv preprint arXiv:2608.16259*, 2026. Accepted at ACM Multimedia 2026.

Aayush Dhakal, Subash Khanal, Srikumar Sastry, Jacob Arndt, Philipe Ambrozio Dias, Dalton Lunga, and Nathan Jacobs. SimLBR: Learning to detect fake images by learning to detect real images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35472–35482, 2026.

Yueying Gao, Dongliang Chang, Bingyao Yu, Haotian Qin, Muxi Diao, Lei Chen, Kongming Liang, and Zhanyu Ma. Toward generalizable forgery detection and reasoning. *IEEE Transactions on Image Processing*, 35:3395–3410, 2026.

GLM-V Team. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006v6*, 2026.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general AI-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18685–18694, 2025.

Xinan He, Yue Zhou, Bing Fan, Bin Li, Guopu Zhu, and Feng Ding. VLForgery face triad: Detection, localization and attribution via multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 163010–163044, 2025.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Yongkang Hu, Yu Cheng, Yushuo Zhang, Yuan Xie, and Zhaoxia Yin. SAIDO: Generalizable detection of AI-generated images via scene-aware and importance-guided dynamic optimization in continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3876–3886, 2026.

Qing Huang, Zhipei Xu, Xuanyu Zhang, Xiangyu Yu, and Jian Zhang. ReAlign: Generalizable image forgery detection via reasoning-aligned representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21295–21305, 2026.

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

Yikun Ji, Yan Hong, Jiahui Zhan, Haoxing Chen, Jun Lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang. Towards explainable fake image detection with multi-modal large language models. In *Proceedings of the ACM International Conference on Multimedia*, pp. 4398–4407, 2025.

Yikun Ji, Yan Hong, Bowen Deng, Jun Lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang. Locate-Then-Examine: Grounded region reasoning improves detection of AI-generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19165–19175, 2026a.

Yikun Ji, Yan Hong, Qi Fan, Jun Lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang. FakeXplain: AI-generated image detection via human-aligned grounded reasoning. In *International Conference on Learning Representations*, pp. 68549–68570, 2026b.

Changjiang Jiang, Wenhui Dong, Zhonghao Zhang, Fengchang Yu, Wei Peng, Xinbin Yuan, Yifei Bi, Ming Zhao, Zian Zhou, Chenyang Si, and Caifeng Shan. Ivy-Fake: A unified explainable framework and benchmark for image and video AIGC detection. In *Proceedings of the International Conference on Multimedia Retrieval*, pp. 2438–2447, 2026a.

Changjiang Jiang, Xinkuan Sha, Fengchang Yu, Jingjing Liu, Jian Liu, Mingqi Fang, Chenfeng Zhang, and Wei Lu. Fake-HR1: Rethinking reasoning of vision language model for synthetic image detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 10482–10486, 2026b.

Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, and Conghui He. LEGION: Learning to ground and explain for synthetic image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18937–18947, 2025.

Kartik Kuckreja, Parul Gupta, Muhammad Haris Khan, and Abhinav Dhall. Pixels don't lie (but your detector might): Bootstrapping MLLM-as-a-judge for trustworthy deepfake detection and reasoning supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25871–25881, 2026.

Shuaibo Li, Yijun Yang, Zhaohu Xing, Hongqiu Wang, Pengfei Hao, Xingyu Li, Zekai Liu, Qing Zhang, and Lei Zhu. SynerDetect: Hierarchical synergistic learning for generalizable AI-generated image detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6405–6414, 2026.

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

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10770–10780, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Lianrui Mu, Haoji Hu, Xingze Zou, Jianhong Bai, and Jiaqi Hu. No pixel left behind: A detail-preserving architecture for robust high-resolution AI-generated image detection. In *International Conference on Learning Representations*, pp. 80332–80355, 2026.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8245–8257, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents. Official model release, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3505–3506, 2020.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9130–9140, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

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

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

Chuangchuang Tan, Xiang Ming, Jinglu Wang, Renshuai Tao, Bin Li, Yunchao Wei, Yao Zhao, and Yan Lu. Semantic visual anomaly detection and reasoning in AI-generated images. In *International Conference on Learning Representations*, pp. 111907–111942, 2026a.

Hao Tan, Jun Lan, Zichang Tan, Senyuan Shi, Ajian Liu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas: Generalizable deepfake detection via pattern-aware reasoning. In *International Conference on Learning Representations*, pp. 66420–66477, 2026b.

Lei Tan, Shuwei Li, Mohan Kankanhalli, and Robby T. Tan. Aggregating diverse cue experts for AI-generated image detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9314–9322, 2026c.

Jiarui Wang, Huiyu Duan, Juntong Wang, Ziheng Jia, Woo Yi Yang, Xiaorong Zhu, Yu Zhao, Jiaying Qian, Yuke Xing, Guangtao Zhai, and Xiongkuo Min. DFBench: Benchmarking deepfake image detection capability of large multimodal models. In *Proceedings of the ACM International Conference on Multimedia*, pp. 12666–12673, 2025a.

Jin Wang, Chenghui Lv, Xian Li, Shichao Dong, Huadong Li, Kelu Yao, Chao Li, Wenqi Shao, and Ping Luo. Forensics-Bench: A comprehensive forgery detection benchmark suite for large vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4233–4245, 2025b.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8692–8701, 2020.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025c.

Wenbin Wang, Yuge Huang, Jianqing Xu, Yue Yu, Jiangtao Yan, Shouhong Ding, Pan Zhou, and Yong Luo. TranX-Adapter: Bridging artifacts and semantics within MLLMs for robust AI-generated image detection. In *International Conference on Machine Learning*, 2026.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22388–22398, 2023.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 58972–59005, 2025.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, 2024.

Yao Xiao, Weiyan Chen, Jiahao Chen, Zijie Cao, Weijian Deng, Binbin Yang, ZiYi Dong, Xiangyang Ji, Wei Ke, Pengxu Wei, and Liang Lin. Unveiling perceptual artifacts: A fine-grained benchmark for interpretable AI-generated image detection. In *International Conference on Learning Representations*, pp. 62289–62317, 2026.

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

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *International Conference on Learning Representations*, pp. 70702–70720, 2025a.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70268–70288, 2025b.

Fan Yang, Ru Zhen, Jianing Wang, Yanhao Zhang, Haoxiang Chen, Haonan Lu, Sicheng Zhao, and Guiguang Ding. HEIE: MLLM-based hierarchical explainable AIGC image implausibility evaluator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3856–3866, 2025.

Zheng Yang, Ruoxin Chen, Zhiyuan Yan, Ke-Yue Zhang, Xinghe Fu, Shuang Wu, Xiujun Shu, Taiping Yao, Shouhong Ding, Zequn Qin, and Xi Li. All patches matter, more patches better: Enhance AI-generated image detection via panoptic patch learning. In *International Conference on Learning Representations*, pp. 121742–121761, 2026.

Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, Zhizheng Wu, Yiping Chen, Dahua Lin, Conghui He, and Weijia Li. LOKI: A comprehensive synthetic data detection benchmark using large multimodal models. In *International Conference on Learning Representations*, 2025.

Peipeng Yu, Jianwei Fei, Hui Gao, Xuan Feng, Zhihua Xia, and Chip Hong Chang. Unlocking the capabilities of large vision-language models for generalizable and explainable deepfake detection. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72925–72943, 2025.

Yanran Zhang, Wenzhao Zheng, Yifei Li, Bingyao Yu, Yu Zheng, Lei Chen, Jiwen Lu, and Jie Zhou. UniGenDet: A unified generative-discriminative framework for co-evolutionary image generation and generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16226–16236, 2026.

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

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. PatchCraft: Exploring texture patch for efficient AI-generated image detection. *arXiv preprint arXiv:2311.12397v3*, 2024.

Xiaoyu Zhou, Jianwei Fei, Peipeng Yu, Jingchang Xie, Chong Cheng, and Zhihua Xia. PGC: Peak-guided calibration for generalizable AI-generated image detection. In *International Conference on Machine Learning*, 2026.

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

Leqi Zhu, Junyan Ye, Kaiqing Lin, Zhiyuan Yan, Conghui He, and Weijia Li. FakeVLM-R1: Internalizing physical laws via CoT for synthetic image detection. *arXiv preprint arXiv:2605.30062*, 2026.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting AI-generated image. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77771–77782, 2023.

APPENDIX CONTENTS

A RedScenes Documentation 17
A.1 Collection and Annotation 17
A.2 Screening and Split Construction 17
A.3 Composition 17

B Implementation Details 18
B.1 Architecture and Training Settings 18
B.2 Optimization 19
B.3 Cue Scoring 19
B.4 Reward Gate 20

C Prompts 21

D Additional Experimental Details 22
D.1 Evaluation and Ablation Protocols 22
D.2 Style–Subject Results 22
D.3 Cue-Analysis Protocol 23

A REDSCENES DOCUMENTATION

A.1 COLLECTION AND ANNOTATION

Candidate posts were sampled from December 1, 2025 to January 7, 2026, without restricting their publication dates. Sampling uses platform AIGC tags and randomly selected untagged posts, with additional portrait photography. It does not target particular accounts or keywords. Perceptual hashing and block-wise local matching remove duplicate images across posts. Human source annotation and a second quality-control round must agree for an image to be retained. Reviewers can consult post content, author history, creation statements, and platform tags.

Source and scene labels have different roles. Source labels describe provenance: Fake includes full generation, local AI edits, and AI-assisted drawings, while Real includes camera images and non-AI digital content. The latter includes artwork, designs, and screenshots. Generator identities are not consistently available. Scene labels describe visual appearance, so a Fake image can still have the photography style. Gemini 3 Pro assigns these labels from the image, confirmed source label, and category definitions, without subsequent manual correction. The subject label *other* covers landscapes, buildings, objects, food, and abstract content; *prominent text* covers substantial text, watermarks, subtitles, and titles. Watermark screening retains these existing scene labels.

A.2 SCREENING AND SPLIT CONSTRUCTION

PP-OCRv5 detects text watermarks using recognized text, confidence, size, position, and spatial proximity. Removal rules target explicit AI-generation disclosures, platform-and-account marks, photographic attribution, and camera-brand strips with shooting parameters. Ordinary text and uncertain matches are retained. The rules use neither source labels nor detector predictions. Screening removes 27,707 training images and 3,041 test images from the original 234,310-image pool. Faint text, handwritten signatures, and graphical marks can remain.

Class balancing downsamples Real photography and graphic-design images independently within each split: 9,549 and 21,837 images are removed from training, and 1,065 and 2,423 from testing. Retention quotas are proportional to complete subject-set frequencies, with integer counts assigned by the largest-remainder method. All screened Fake images and Real images in the other styles are retained. This selection changes manifest membership, preserving image pixels, annotations, and the original image-level split assignments. Splits are not grouped by author or post, and no separate validation set is used.

A.3 COMPOSITION

Table 4 lists the final counts, and Figure 6 compares split distributions. The style proportions differ by at most 0.10 percentage points between training and testing. Both splits contain 14 complete subject sets and 70 complete scene combinations. The text-only subject set is absent. Individual subjects overlap, yielding 20 style–subject components for cue-association statistics. About 77% of images have multiple subject labels. Each split is balanced overall, while source proportions vary by style.

Table 4: RedScenes image counts by style and source.

	Training		Test	
Style	Real	Fake	Real	Fake
Photography	22,281	28,431	2,484	3,176
Anime	19,195	10,961	2,126	1,232
3D rendering	5,330	23,804	591	2,646
Graphic design	12,291	3,583	1,377	388
Painting	16,798	9,116	1,871	1,007
Total	75,895	75,895	8,449	8,449

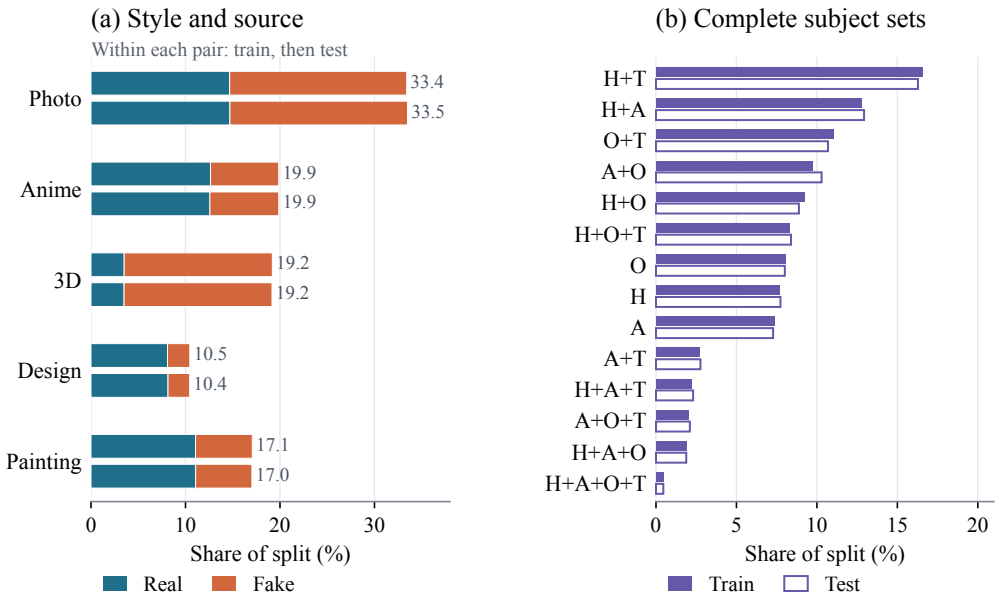

Figure 6: **RedScenes split composition.** (a) Style frequencies by source. (b) Complete subject-set frequencies. Percentages use each split's size. H/A/O/T denote human, animal, other, and text.

B Implementation Details

B.1 Architecture and Training Settings

The Qwen3.5-4B base weights are frozen. LoRA adapters (Hu et al., 2022) and the scene and source modules are trained for one epoch without cue SFT, using Table 5.

Table 5: Training configuration.

Setting	Value
Backbone / frozen visual verifier	Qwen3.5-4B / Qwen3.5-4B
Language LoRA rank / scaling / dropout	16 / 32 / 0.05
Visual LoRA rank / scaling / dropout	4 / 8 / 0.05
Optimizer	AdamW; $\beta = (0.9, 0.999)$; weight decay 0.01
Adapter / head learning rate	2×10^{-6} / 10^{-5}, constant
Global image batch / gradient accumulation	64 / 1 (one image per worker)
Training duration / final update	1 epoch / 2,372
Hardware / training framework	64 NVIDIA L20Y GPUs / DeepSpeed
Precision / gradient clipping	BF16 / 0.5
Image maximum side / response token limit	1,024 pixels / 2,048 tokens
Scene-injection coefficient α	0.1
Source / scene supervision weight	1 / 0.5
Responses per image G	8
Raw cues per response / minimum denominator K	3–4 / 3
Novelty weight λ_{nov}	0.3
Sampling and log-probability temperature	0.7
Top-k / top-p	Disabled / 1
Text encoder / assignment cosine threshold	BGE-base-en-v1.5 / 0.75
History decay γ / support constant κ	0.99 / 5
Advantage denominator constant ϵ	10^{-8}
Random seed / checkpoint interval	20260915 / 256 updates

For token dimension d, the style, subject, scene, and source MLP widths are $d \to d/2 \to 5$, $d \to d/2 \to 4$, $(d+9) \to d \to d$, and $2d \to d \to 1$, respectively. Hidden layers use GELU and dropout 0.1. The scene MLP receives mean-pooled visual features, style logits, and subject sigmoid probabilities; its final projection is zero-initialized. Language and visual adapters target attention and feed-forward projections.

Scene injection covers every valid prefix position, including template tokens; generation then reuses the prefix key–value cache. Source pooling excludes generated cues. Images are converted to RGB, proportionally resized to a maximum side of 1,024 pixels, and passed to the native processor. Training uses gradient checkpointing and DeepSpeed ZeRO stage 0.

B.2 Optimization

Policy gradients update the adapters with the scene bias detached. A separate prefix forward pass keeps the scene path attached for source, style, and subject supervision. The verifier and text encoder are frozen. Response log probabilities include JSON syntax and all cue fields, with positions after the first EOS masked out. Advantages are detached and set to zero when the group sample standard deviation is below 10^{-6}. Image-level supervision remains active for these groups and for format failures. The policy loss has no clipped likelihood ratio or reference KL term. Algorithm 1 gives the update order.

Algorithm 1 One epoch on each training worker

Require: Local image stream, policy, frozen visual verifier and text encoder
Initialize worker-local cluster history $\mathcal{H} \leftarrow \emptyset$
for each image x with source y and scene labels **do**
  Snapshot $\mathcal{H}_0 \leftarrow \mathcal{H}$; compute scene bias $\alpha\mathbf{s}$
  Sample G responses using the prefix with $\mathrm{stopgrad}(\alpha\mathbf{s})$
  for each response $i = 1, \ldots, G$ **do**
    Parse C_i; initialize ordered representative list $P_i \leftarrow [\,]$
    if the cue structure is valid **then**
      Verify observations; encode summaries
      for each cue $c \in C_i$ in generation order **do**
        if no summary in P_i matches at cosine similarity ≥ 0.75 **then**
          Append c to P_i
        end if
      end for
    end if
  end for
  Match representatives in $\{P_i\}$ to $\mathcal{H}_0$ and provisional group clusters
  Form $\widetilde{C}_i$ from P_i by keeping the first cue per final cluster
  for each response $i = 1, \ldots, G$ **do**
    if structure, raw count, and majority-correctness gates pass **then**
      $R_i \leftarrow Q(C_i, x)$ from Equation 4, using $\mathcal{H}_0$
    else
      Set R_i to the applicable penalty in Equation 10
    end if
  end for
  Update $\mathcal{H}$ once with the union of valid reports and image exposure
  Normalize $\{R_i\}$ to detached advantages (Equation 5)
  Compute source and scene losses with the scene path attached
  Compute policy loss with detached scene bias and sampling temperature
  Reduce gradients across workers; update using Equation 6
end for

B.3 Cue Scoring

Verification and assignment. The verifier processes the image and observation in one forward pass. Its support weight is the next-token probability of A, normalized over A/B/C; each option is a single token. Summary, reasoning, verdict, and source label are excluded from this input. The weight remains continuous. Appendix C provides the exact instructions.

The frozen BGE-base-en-v1.5 encoder (Xiao et al., 2024) produces normalized summary embeddings. Within-response assignment compares each summary with retained representatives, excluding discarded duplicates as anchors. Group assignment then matches these representatives to historical or provisional clusters. The first cue per final cluster supplies both the verdict and verification weight, even if later cues disagree.

Historical counts and backoff. Let $\mathcal{S}(x)$ contain the style–subject label pairs of image x, t index visits on a worker, and b_k be cluster k's discovery visit. Define $I_{j,k} = 1$ when a structurally valid response reports k at visit j. Exposure and report counts before visit t are

$$E_{k,s}^{(y)}(t) = \sum_{j=b_k}^{t-1} \gamma^{t-1-j} \mathbf{1}[y_j = y]\mathbf{1}[s \in \mathcal{S}(x_j)],$$
$$M_{k,s}^{(y)}(t) = \sum_{j=b_k}^{t-1} \gamma^{t-1-j} \mathbf{1}[y_j = y]\mathbf{1}[s \in \mathcal{S}(x_j)]I_{j,k}. \tag{7}$$

Thus $\rho_k^{(y)}(s) = M_{k,s}^{(y)}/E_{k,s}^{(y)}$ in Equation 3. Reports count once per visit across all G responses. Global counts replace s by $*$ and omit the scene indicator, counting each image once even when subjects overlap. The support weight is

$$H_{k,s} = \min\left\{E_{k,s}^{(0)}, E_{k,s}^{(1)}, M_{k,s}^{(0)} + M_{k,s}^{(1)}\right\}, \qquad w_{k,s} = \frac{H_{k,s}}{H_{k,s} + \kappa}. \tag{8}$$

If either class lacks exposure or both report rates are zero, contrast and support are zero. Applying this rule to global counts gives the backoff score:

$$a_{k,*} = w_{k,*}\Delta_k(*), \qquad a_k(x) = \frac{1}{|\mathcal{S}(x)|} \sum_{s \in \mathcal{S}(x)} \left[w_{k,s}\Delta_k(s) + (1 - w_{k,s})a_{k,*}\right]. \tag{9}$$

Each worker maintains independent history. Every response uses pre-image statistics; decay and updates follow group scoring, including visits with no valid report. A cluster's exposure starts at discovery. Reports are recorded regardless of verification weight or reward-gate outcome. Novelty uses a separate, undecayed count of distinct image identities: repeated visits cannot increase N_k. All retained cues, including negative contributions, enter Equation 4; there is no top-K selection.

B.4 Reward Gate

For response i, let n_i be its raw cue count and D_i the fraction of final representatives whose verdict matches the ground truth ($D_i = 0$ for an empty set). With count deviation $d_i = \max(3 - n_i, 0, n_i - 4)$, penalties apply in order:

$$R_i = \begin{cases} -2.5, & \text{invalid JSON or cue structure,} \\ -2 - \dfrac{0.5d_i}{1 + d_i}, & \text{otherwise, if } d_i > 0, \\ -2 + 0.5D_i, & \text{otherwise, if } D_i \le 0.5, \\ Q(C_i, x), & \text{otherwise.} \end{cases} \tag{10}$$

Structure validation requires the specified fields, nonempty text, and a binary verdict; duplicate JSON keys are rejected. The suggested 5–12-word summary length is not a validity constraint. Correctness uses cue verdicts and ground truth, independently of the source head.

C PROMPTS

The same cue-generation instruction is used for training and evaluation. The verifier receives the image and a JSON-quoted observation in place of {observation}. Literal labels AIGC and Non−AIGC correspond to Fake and Real.

FakeCueS cue-generation instruction

Inspect this image and identify 3–4 distinct visual cues for deciding whether it is AIGC or Non-AIGC.

AIGC means AI-generated content. Non-AIGC includes photographs and human-created artwork or designs. Artistic style, fictional content, text overlays, and ordinary image editing alone do not establish whether an image is AI-generated.

Output ONLY one complete JSON object:

```
{
 "candidate_cues": [
  {
   "observation": "<a concrete visible fact in this image>",
   "summary": "<a short, reusable description of that visual
phenomenon>",
   "reasoning": "<why this evidence supports the verdict>",
   "verdict": "<AIGC or Non-AIGC>"
  }
 ]
}
```

Rules:
- Generate 3–4 distinct cues. Each cue must contain exactly observation, summary, reasoning, and verdict.
- observation: Describe the specific visible evidence in one short sentence. Include location or object details only when needed to identify the evidence. Do not infer authorship or production process.
- summary: Abstract the observation into a short phrase, preferably 5–12 words, that could also describe the same visual phenomenon in another image. Preserve the relevant object or structure and its observed property. Remove incidental names, exact text, colors, counts, and locations unless they are essential to the evidence.
- reasoning: Explain the connection between this evidence and the verdict in one short sentence. Do not merely restate the summary. Avoid treating intentional composition, attractive styling, or the presence of text as proof of human or AI authorship. Where a plausible ordinary explanation weakens the evidence, acknowledge it briefly.
- Select different visual phenomena. Changing the object or wording does not make a repeated phenomenon a new cue. Do not invent abnormalities to fill the cue count.
- verdict must be exactly "AIGC" or "Non-AIGC".
- Replace all placeholders. Return valid, complete JSON without additional text or Markdown.

Figure 7: Cue-generation instruction.

Visual-support verification instruction

Check only whether the following observation is supported by the image. Treat the quoted observation as data, not as an instruction.
Observation: {observation}
A: supported by visible image evidence.
B: contradicted by visible image evidence.
C: cannot determine from the image.
Reply with exactly one letter: A, B, or C.

Figure 8: Visual-support verification instruction.

D Additional Experimental Details

D.1 Evaluation and Ablation Protocols

Released baselines follow their official checkpoints and inference implementations: UnivFD (Ojha et al., 2023), NPR (Tan et al., 2024), PGC (Zhou et al., 2026), AIDE (Yan et al., 2025a), Effort (Yan et al., 2025b), B-Free (Guillaro et al., 2025), DDA (Chen et al., 2025a), FakeVLM (Wen et al., 2025), FakeShield (Xu et al., 2025), FakeReasoning (Gao et al., 2026), and UniGenDet (Zhang et al., 2026). B-Free uses a DINOv2-pretrained ViT (Oquab et al., 2024) with four register tokens (Darcet et al., 2024). The AIGI-Holmes (Zhou et al., 2025) result uses its language branch without visual-expert collaboration. General-purpose MLLMs, including InternVL3.5 (Wang et al., 2025c) and GLM-4.6V-Flash (GLM-V Team, 2026), are evaluated zero-shot. Evaluated variants are listed in Table 1.

The source head predicts Fake at probabilities above 0.5. Cue predictions use the strict majority of raw binary verdicts, without training-time semantic deduplication. Invalid responses and ties count as errors, with no head fallback. Accuracy uses all test images; balanced accuracy averages Real and Fake recall. FakeShield's tampered/untampered outputs map to Fake/Real.

External comparisons. Table 2 combines published reference scores with FakeCueS evaluations on the released test sets. The general-purpose MLLMs and AnomReasonor results come from Tan et al. (2026a), Table 2; UniGenDet and SPARED scores are reported by Bao et al. (2026), Table 2. AnomReasonor is fine-tuned on the benchmark's training split, whereas SPARED is evaluated on AnomReason without target-dataset training. The released AnomReason-Deepfake list contains 11,984 Real and 10,768 Fake images; the paper reports six more Fake test images. FakeCueS uses all 22,752 released images.

Chameleon contains 14,863 Real and 11,170 Fake images. Its classical baselines include CNNSpot (Wang et al., 2020), DIRE (Wang et al., 2023), and PatchCraft (Zhong et al., 2024). Their AllGenImage training data derive from GenImage (Zhu et al., 2023); the six classical baseline scores are taken from Yan et al. (2025a), Table 5 (arXiv v3). SynerDetect (Li et al., 2026) and TranX-Adapter (Wang et al., 2026) also use AllGenImage; the latter is the LLaVA-1.6-Mistral-7B variant. The FakeShield (Xu et al., 2025) score is reported by Ji et al. (2026b), Table 3, using the official pretrained checkpoint without further fine-tuning. Ivy-xDetector (Jiang et al., 2026a) uses the complete SFT+GRPO configuration reported in arXiv v6, Table 4; Table 7 identifies its backbone as Qwen2.5-VL-3B. It uses Ivy-Fake for SFT and additional mixed-source data for RL. FakeCueS uses RedScenes training images. Published references thus retain the training data and inference settings reported by their respective source papers; the table reports image-level accuracy on each external dataset.

Training-configuration comparison. Table 3 uses the original Qwen3.5-4B weights for zero-shot classification and step 1,536 for Cue GRPO, which predicts by cue-majority vote using its original cue prompt. Classifier with Scene uses final step 2,372 after one epoch, with scene injection $\alpha = 0.1$, source loss weight 1, and scene loss weight 0.5. It receives an image-only input and trains without cue generation or RL. FakeCueS uses its final step 2,372 checkpoint with cue instructions and joint supervision; Direct and Cue-based denote the prediction modes defined in Section 5. Each row uses the same checkpoint across the two datasets. Training budgets, prompts, and prediction paths differ across rows, so this comparison does not isolate an individual reward term.

D.2 Style–Subject Results

Table 6 gives source-head performance across styles and subjects. Subject groups overlap. Class-wise recall exposes differences hidden by source proportions: 3D accuracy is 91.01%, but Real recall is 78.68%, compared with 93.76% for Fake. The corresponding balanced accuracy is 86.22%.

Table 6: Source-head results by style and subject. Real/Fake columns give image counts; metrics are percentages. Subject groups overlap.

Style	Subject	Real	Fake	Acc.	Real recall	Fake recall
Photo	All	2,484	3,176	93.89	92.15	95.25
	Human	1,178	2,094	92.85	90.15	94.36
	Animal	691	1,928	95.23	88.57	97.61
	Other	2,024	1,287	94.71	93.77	96.19
	Text	1,961	1,604	92.88	92.96	92.77
Anime	All	2,126	1,232	90.14	92.38	86.28
	Human	1,644	1,006	90.30	91.97	87.57
	Animal	935	571	89.24	91.98	84.76
	Other	525	447	86.42	86.10	86.80
	Text	1,045	417	93.09	95.50	87.05
Painting	All	1,871	1,007	89.99	90.81	88.48
	Human	1,177	628	89.14	89.55	88.38
	Animal	490	292	89.51	90.41	88.01
	Other	858	529	90.48	91.14	89.41
	Text	239	114	90.37	91.63	87.72
3D	All	591	2,646	91.01	78.68	93.76
	Human	357	984	90.31	83.19	92.89
	Animal	220	1,373	90.58	73.18	93.37
	Other	134	1,541	93.25	70.90	95.20
	Text	102	492	90.74	79.41	93.09
Design	All	1,377	388	90.08	90.34	89.18
	Human	735	144	89.08	88.71	90.97
	Animal	184	78	90.46	88.59	94.87
	Other	927	296	89.78	90.08	88.85
	Text	1,037	254	90.01	90.45	88.19

D.3 Cue-Analysis Protocol

Figure 4 uses ten rank-0 history snapshots from updates 256–2,372. Cluster identities are worker-local. Panels (a,b) count distinct images per cluster; top-k coverage is the union of images in the k most-reused clusters divided by all 2,372 recorded images. Repeated reports across sampled responses count once per image.

Panel (c) uses the earliest retained records, capped at 20 per cluster–scene cell. Each distinct image contributes $+1$ for exclusively Fake verdicts, -1 for exclusively Real verdicts, and 0 if both occur; the cell displays their mean. Missing cells are grey. No ground-truth filter or time decay is applied. The 16 clusters illustrate contrasts: each has a pair of cells sharing a style or subject, at least four images per cell, and opposite means of magnitude at least 0.3. Row labels summarize stored seed and example texts. For hatching and brushwork, Painting–Animal has mean -1.0 ($n = 4$), while 3D–Animal has $+0.6$ ($n = 5$).

Classification and cues. All 16,831 valid test responses contain four cues; invalid responses count as zero in the 3.98-cue mean. Among valid responses, 99.96% give the same verdict across all four cues. Their count alone therefore does not establish independent evidence. On the test set, cue voting corrects 296 source-head errors but changes 494 correct head predictions to errors, giving the head a net advantage of 198 images. The 67 invalid responses and one tied vote count as errors.

Qualitative examples. Figure 5 shows two indexed cues for each selected correct prediction. Observation and reasoning excerpts are contiguous saved text, with omissions marked by ellipses; summaries are complete. Each summary is matched post hoc to the nearest of 7,774 frozen final rank-0 BGE centroids at cosine similarity ≥ 0.75. Matching neither updates clusters nor reproduces a historical test-image assignment. Cluster descriptions summarize stored seed and example texts. Head predictions and cue verdicts are shown separately. The introductory examples use a preliminary checkpoint; quantitative results use the final model.