UC-VLM: Consistency-Driven Learning for AI-Generated Image Detection with Vision-Language Large Models

Lei Tan[1], Shuwei Li[1], Mohan Kankanhalli[1], and Robby T. Tan[1]

National University of Singapore, Singapore
lei.tan@nus.edu.sg, shuwei@u.nus.edu, mohan@comp.nus.edu.sg, robby.tan@nus.edu.sg

Abstract. Vision-Language Large Models (VLLMs) are promising for AI-generated image (AIGI) detection because they can produce both a prediction and a natural-language output. However, most existing VLLM-based detectors primarily fine-tune the language side while giving limited attention to low-level visual forensic cues. They also often depend on manually crafted prompts or human-annotated rationales, which limits its scalability. We present **UC-VLM**, a unified multi-stage framework for AIGI detection that relies solely on binary supervision. UC-VLM first identifies effective instruction variants automatically. It then reuses the same binary label within a multi-stage training framework: (i) a visual discrimination objective that strengthens sensitivity to non-semantic forensic cues, and (ii) a label-conditioned generation objective that uses the binary label to supervise textual outputs. This design turns weak binary supervision into a shared supervision signal for both the visual pathway and the language output. Our key novelty is a unified multi-stage binary-supervised framework that consistently reuses the same authenticity labels for visual adaptation and label-conditioned text generation, while leveraging automatically optimized instructions to reduce prompt sensitivity without requiring human-written rationales or handcrafted prompts. Experiments show that UC-VLM achieves **96.1%** average accuracy on GenImage, exceeding the strongest prior result by **4.6%**, and obtains **69.6%** / **77.9%** accuracy on Chameleon under ProGAN / SDV1.4 training, surpassing the best baseline by **11.2%** / **15.3%**, respectively.

Keywords: AI-Generated Image Detection, Vision-Language Models

1 Introduction

Recent advances in generative AI have made it possible to synthesize highly realistic images that can convincingly mimic natural photographs [27, 38, 44, 51, 55]. These capabilities are transforming content creation, but they also intensify concerns about the authenticity and trustworthiness of digital media [13, 23–26].

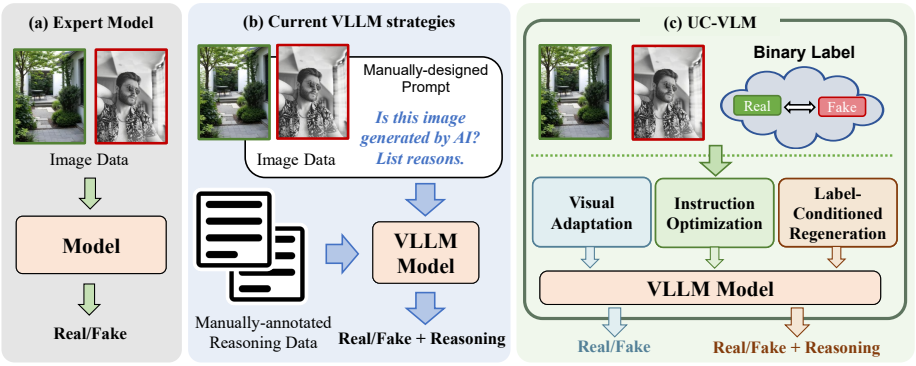

Fig. 1: Comparison of paradigms for AI-generated image detection. (a) Conventional classifiers predict real/fake from image data. (b) Existing VLLM-based detectors often rely on manually crafted prompts and human-annotated rationales to produce a label with accompanying text. (c) UC-VLM learns from binary authentic/generated labels only, combining visual adaptation, instruction optimization, and label-conditioned generation to improve robustness to prompt choice and provide scalable supervision for textual outputs.

Most existing AI-generated image (AIGI) detectors formulate the task as binary classification [50,53]. As illustrated in Figure 1(a), models are trained on datasets containing both real and synthesized images and learn a discriminative boundary between the two classes. To improve separability, prior work often augments the input with explicit transformations or auxiliary signals [34, 53, 59], leveraging cues such as neighboring-pixel statistics [53], high-frequency artifacts [52], or reconstruction-based discrepancies [48]. While these strategies can yield strong benchmark accuracy, they do not naturally support textual outputs and often show limited robustness beyond their training conditions.

With recent advances in Vision-Language Large Models (VLLMs) (e.g., [19, 29, 58, 66]), research has begun to explore their potential for AIGI detection [15, 60, 71]. Unlike traditional classifiers, VLLMs align visual and textual representations through large-scale multimodal pretraining, enabling a unified interface for prediction and text generation. Despite this promise, current VLLM-based detectors [15, 60] still face key limitations, as illustrated in Figure 1(b). Prior successful detectors have shown that the discriminative differences between real and generated images often lie in subtle low-level details, such as frequency artifacts [52], local textures [53], and non-semantic forensic traces [10]. They often prioritize language-side fine-tuning while giving limited attention to low-level forensic cues in the visual pathway. They also rely on manually crafted prompts and human-annotated rationales, which limits scalability. Moreover, such handcrafted designs underexplore how prompt formulation affects prediction stability and detection consistency.

To address these challenges, we introduce **UC-VLM**, a unified learning framework for AIGI detection, as illustrated in Figure 1(c). UC-VLM is trained with only authentic/generated labels, but it treats these binary labels as more

than final decision targets. Our key idea is to expand weak binary supervision into a shared training signal that is consistently reused across visual adaptation, instruction optimization, and language-side generation in a unified multi-stage framework. This shared-supervision design is important because, in VLLM-based detection, prediction stability is jointly affected by (i) what evidence the vision encoder exposes, (ii) how the instruction queries that evidence, and (iii) how the language model expresses the output under the same supervision.

Optimizing these aspects in isolation can be unreliable: visual tuning alone may overfit to low-level patterns, prompt selection alone can lead to unstable behavior under rephrasing, and language-side tuning alone may improve text fluency without improving the underlying visual evidence used for detection. UC-VLM addresses this by reusing the same binary labels to enforce three tightly connected behaviors within a single learning framework. First, UC-VLM adapts the vision encoder to increase sensitivity to local non-semantic forensic cues, encouraging the model to rely less on high-level semantics and more on artifact-level evidence. Second, UC-VLM performs instruction optimization to automatically discover effective instruction formulations for authenticity assessment. By selecting the best-performing instruction from an automatically generated pool, the model obtains a more stable linguistic interface. Third, UC-VLM applies label-conditioned text generation so that the binary label also provides supervision for language-side outputs, enabling scalable textual-output training without human-annotated rationales. Together, these behaviors turn binary authenticity labels into multi-stage optimization on visual discrimination, prompt robustness, and textual-output generation, under a shared supervision signal.

Concretely, UC-VLM is implemented as a unified multi-stage training procedure. It first adapts the vision encoder with a visual discrimination loss, and then refines the language module with a label-conditioned generation loss while keeping the vision tower fixed. Although these stages are optimized sequentially, they are unified by the same binary authenticity supervision: the visual stage shapes the evidence available to the model, while the generation stage transfers the same supervision to language-side outputs. In this sense, UC-VLM is not a set of independently supervised modules, but a multi-stage binary-supervised framework that reuses the same labels across visual and language components. Overall, our contributions are summarized as follows:

- We propose **UC-VLM**, a unified learning framework for VLLM-based AI-generated image detection using only binary authentic/generated labels. UC-VLM combines visual adaptation, instruction optimization, and label-conditioned generation within a unified multi-stage framework, enabling optimization of the visual pathway and language-side outputs under the same supervision signal.
- We show that unified binary-supervised learning improves detection stability beyond standard accuracy. UC-VLM automatically discovers effective instruction formulations and enables scalable textual-output training under binary supervision, without manual prompts and rationales.

– We conduct extensive experiments on GenImage and Chameleon, showing that UC-VLM improves accuracy over the prior results by +**4.6%** on GenImage and +**11.2%** / +**15.3%** on Chameleon under ProGAN / SDV1.4 training, and yields more stable predictions under instruction variants.

2 Related Work

AI-generated image detection has become increasingly important with the rapid progress of modern generative models. Existing methods can be broadly grouped into three lines: hand-crafted forensic cues, deep-learning-based visual detectors, and recent VLM-based detectors.

Early forensic approaches mainly relied on manually designed low-level cues, such as chromatic aberration [39], color saturation [40], blending boundaries [22], and reflection inconsistencies [42]. These methods provide interpretable evidence for specific types of manipulation or synthesis artifacts. However, as generative models become increasingly diverse and photorealistic, such hand-crafted cues are often brittle and may fail to generalize across different generators, resolutions, or post-processing conditions.

With the success of deep learning [16, 35–37, 46, 69], AIGI detectors [31, 32, 57, 68] have achieved strong in-distribution performance. CNNSpot [57] shows that standard convolutional architectures can capture discriminative synthesis traces, while FreDect [14] exploits frequency-domain representations to reveal artifacts that are less visible in the spatial domain. As diffusion models improve perceptual quality and weaken obvious low-level cues, recent methods have explored subtler discriminative cues to improve generalization [61, 65]. DIRE [59] leverages reconstruction discrepancies to distinguish real and generated images, while localized and region-level representations have also been explored to capture fine-grained visual structures, which have been further adopted by other works [34, 67]. ZED [10] identifies discrepancies between real and generated images via coding differences from a lossless encoder in a multi-resolution structure. FatFormer [30], Effort [63], and MCAN [54] improve CLIP-based detectors through frequency-domain fine-tuning, orthogonal-space adaptation, and multi-cue aggregation, respectively. NPR [53] revisits upsampling artifacts to amplify subtle synthesis cues, while AIDE [62] and DEUA [18] enhance detection with multi-frequency learning and uncertainty estimation. Despite these advances, most deep detectors remain purely visual. They usually treat binary labels only as classification targets and provide limited language-side reasoning, making their decisions less transparent and less flexible for VLM-based detection scenarios.

Recently, large Vision-Language Models (VLMs) have motivated a new direction for AIGI detection by enabling language-conditioned prediction and textual responses. FakeVLM [60] fine-tunes the language component of a VLM on the well-annotated FakeClue dataset to produce authenticity judgments with accompanying text. FakeReasoning [15] employs dual visual encoders by combining CLIP and DINO features, providing complementary visual cues for language-

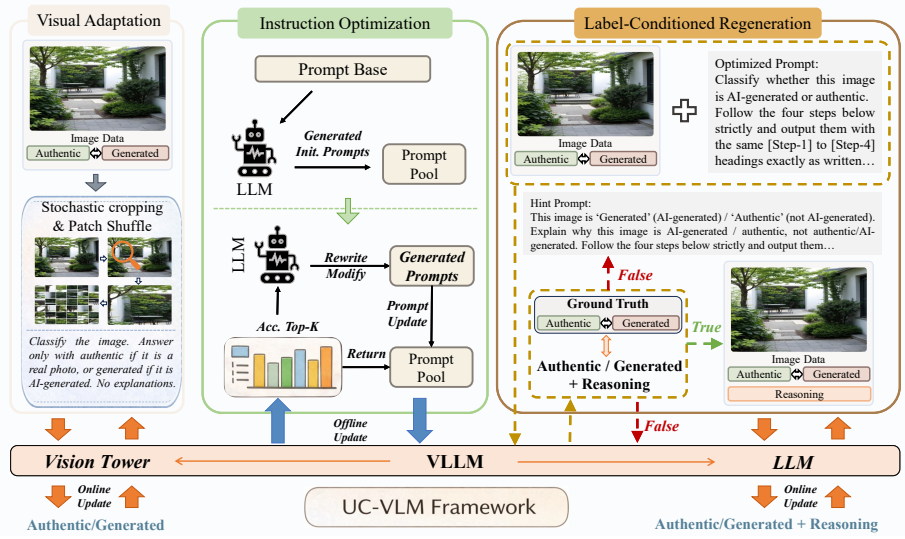

Fig. 2: Overview of UC-VLM. UC-VLM is built upon Qwen2.5-VL-7B and learns AI-generated image detection under binary authentic/generated supervision. The framework includes three coordinated components: (i) **Visual Adaptation**, which enhances the vision encoder's sensitivity to local non-semantic forensic cues; (ii) **Instruction Optimization**, which automatically searches for effective instruction variants for authenticity assessment; and (iii) **Label-Conditioned Regeneration**, which conditions text generation on the binary label to provide scalable supervision for textual outputs while freezing the vision tower.

side generation. AIGI-Holmes [71] integrates CLIP and NPR representations and adopts human-aligned Direct Preference Optimization to improve output consistency. These methods demonstrate the promise of VLM-based detectors, but they often rely on manually curated VQA-style prompts, human-written explanations, or annotated rationales. Such requirements increase annotation cost and may limit scalability when moving to new datasets or generators.

Different from these prior directions, UC-VLM aims to learn a VLM-based detector using only binary supervision. Instead of requiring human-written rationales, UC-VLM consistently reuses the same real/fake labels across visual adaptation, instruction optimization, and label-conditioned language refinement. This design enables binary supervision to guide both the visual pathway and language-side output generation in a scalable framework.

3 Proposed Method

The overall framework of UC-VLM is illustrated in Figure 2. Built upon a vision-language backbone (Qwen2.5-VL-7B) [7], UC-VLM addresses AIGI detection under binary authenticity supervision. Given an input image I, the objective is to predict its authenticity label $y \in \{\text{Authentic}, \text{Generated}\}$, indicating whether the image is captured from the real world or synthesized by a generative model.

During training, only image-level binary labels are available, without human-written rationales, manipulation masks, or fine-grained forensic annotations. Given an image-label pair (I, y), UC-VLM reuses the binary authenticity label as the common supervision source across multiple training stages. Rather than using the binary label only as a final classification target, UC-VLM expands this weak supervision into three complementary aspects of learning. Specifically, the framework includes (i) a visual adaptation stage that strengthens the vision encoder's sensitivity to non-semantic forensic cues, (ii) an instruction optimization component that automatically identifies an effective instruction formulation for authenticity assessment, and (iii) a label-conditioned generation stage that transfers the same binary supervision signal to language-side outputs while freezing the vision tower. In this way, non-semantic forensic cues, prompt robustness, and textual-output training are consistently guided by the same binary label, rather than requiring extra sources of supervision. The training objectives used across the stages of UC-VLM can be summarized as:

$$\mathcal{L} = \mathcal{L}_{vis} + \mathcal{L}_{text}, \tag{1}$$

where $\mathcal{L}_{vis}$ is used for visual-pathway adaptation and $\mathcal{L}_{text}$ is used for label-conditioned text generation. In practice, these objectives are instantiated in a multi-stage training procedure, with visual adaptation performed before language-side generation. Thus, UC-VLM provides a unified binary-supervised framework by consistently reusing the same authenticity labels across visual adaptation, instruction optimization, and language-side generation, rather than introducing additional rationale annotations.

3.1 Visual Adaptation

Prior studies on AIGI detection [34,53] show that authenticity discrimination often depends on subtle low-level artifacts, such as texture inconsistencies, boundary distortions, and frequency irregularities. However, the visual backbone of a VLLM is typically optimized for general semantic understanding, which may underemphasize the local non-semantic evidence needed for forensic discrimination. To improve sensitivity to such cues, we introduce a visual adaptation mechanism that suppresses reliance on global semantics while preserving local structural statistics.

Specifically, given an input image $I \in \mathbb{R}^{H\times W\times 3}$, we first sample a local crop $I' \in \mathbb{R}^{h'\times w'\times 3}$, where

$$h' \sim U\left(\tfrac{1}{8}H, \tfrac{1}{4}H\right), \quad w' \sim U\left(\tfrac{1}{8}W, \tfrac{1}{4}W\right). \tag{2}$$

This stochastic cropping limits access to full-scene semantics while retaining fine-grained textures. As a result, the model is encouraged to rely less on object identity or scene composition and more on local evidence relevant to authenticity.

Following the patch-based processing of Qwen2.5-VL-7B, the cropped image is divided into N non-overlapping patches $\{p_i\}_{i=1}^{N}$. We further randomly permute

Table 1: Effect of prompt formulation on the Chameleon dataset using the pretrained Qwen2.5-VL-7B model. Structured prompts (e.g., P-3) yield higher detection accuracy, indicating that VLM performance is strongly influenced by prompt formulation.

Prompt ID	Accuracy (%)
P-1: *Authentic/Generated QA*	55.5
P-2: *Real/Fake QA*	59.8
P-3: *Answer with Explanation*	63.6

their spatial order:

$$\tilde{I}' = \text{Reassemble}\left(\{p_{\pi(i)}\}_{i=1}^{N}\right), \quad \pi \sim \mathcal{P}(N). \tag{3}$$

This patch shuffling disrupts global spatial coherence while preserving local statistics. Consequently, the model is discouraged from depending on high-level semantic layout and instead pushed toward textural and structural irregularities that are more relevant for distinguishing authentic from generated images.

The transformed image is then encoded into visual embeddings $\{\mathbf{v}_i\}_{i=1}^{N}$, which are fused with textual tokens to produce a binary authenticity prediction:

$$\hat{y} = f_{\text{VLM}}\left(\{\mathbf{v}_i\}_{i=1}^{N}, \mathbf{t}\right). \tag{4}$$

We optimize the visual adaptation term using binary cross-entropy,

$$\mathcal{L}_{\text{vis}} = -\left[y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right]. \tag{5}$$

Under this objective, the vision pathway is directly tuned toward artifact-aware representations using the same authenticity supervision signal that also drives the rest of UC-VLM. This design is important for the overall framework: the visual term does not act as a standalone pretraining step, but as one component of the shared binary-supervised learning process that supports instruction robustness and language-side refinement.

3.2 Instruction Optimization

The formulation of instructions can significantly influence the behavior of VLLMs. Here, we use the term *instruction* to denote the textual query used for authenticity assessment. To illustrate this sensitivity, we evaluate the backbone model (Qwen2.5-VL-7B) on the Chameleon dataset under different instruction formulations:

P-1: `Classify if this image is AI-generated or authentic. Answer only with ‘‘Authentic’’ or ‘‘Generated’’. No explanations.`

P-2: `Classify if this image is real or fake. Answer only with ‘‘Real’’ or ‘‘Fake’’. No explanations.`

P-3: `Evaluate the image and decide if it is ‘‘Generated’’ (AI-generated) or ‘‘Authentic’’ (not AI-generated). Follow the four steps below.`

Algorithm 1: Instruction Optimization

Input: LLM, VLM, BASE_PROMPT, pool size N, Top-k ($k < N$), rounds M
Output: Prompt-B, Instruction Pool
Initialize Instruction Pool $\mathcal{P}_0 = \{p_i\}_{i=1}^{N}$ generated by LLM from BASE_PROMPT
for $t = 0$ **to** $M - 1$ **do**
  Evaluate each $p \in \mathcal{P}_t$ on VLM, obtain $\mathrm{Acc}(p)$
  Select $\mathcal{P}_t^* = \text{Top-}k(\mathcal{P}_t)$ by $\mathrm{Acc}(p)$, set Prompt-B $= \arg\max_{p \in \mathcal{P}_t^*} \mathrm{Acc}(p)$
  Sample Random Parent Instructions $\mathcal{R}_t \subseteq \mathcal{P}_t^*$
  Generate Refined Instructions $\tilde{\mathcal{P}}_t$ using LLM:
  $p' = \begin{cases} \mathrm{Rewrite}(p), & \text{with prob. } 0.5 \\ \mathrm{Modify}(p), & \text{with prob. } 0.5 \end{cases}, \forall p \in \mathcal{R}_t$
  Update Instruction Pool: $\mathcal{P}_{t+1} = \mathcal{P}_t^* \cup \tilde{\mathcal{P}}_t$, $|\mathcal{P}_{t+1}| = N$
return Prompt-B, Instruction Pool

- `Step-1: Visual Examination: List up to three suspicious and three realistic visual cues (e.g., glitches, lighting, motion blur).`
- `Step-2: Statistical Analysis: Highlight frequency-domain evidence supporting or contradicting authenticity.`
- `Step-3: Semantic & Physics Consistency: Identify impossible reflections, shadow errors, or physical mismatches.`
- `Step-4: Final Assessment: Give one word stating ‘‘Generated’’ or ‘‘Authentic’’, and summarize key evidence.`

As shown in Table 1, even minor lexical variations can noticeably affect detection accuracy. Replacing ‘Authentic/Generated’ with ‘Real/Fake’ already yields a measurable improvement, while adding structured guidance further improves performance. These results indicate that the backbone VLLM already contains useful knowledge for authenticity discrimination, but its predictions remain sensitive to the linguistic form of the instruction.

Motivated by this sensitivity, we introduce an automatic instruction optimization procedure to identify instruction variants that are both effective and relatively stable for AIGI detection. Rather than relying on manually designed prompts, we construct an instruction pool and iteratively refine it through generation, evaluation, and transformation, as illustrated in Figure 2 and Algorithm 1.

Starting from a structured base template (*BASE_ PROMPT*), the LLM generates N candidate instructions to initialize an instruction pool. Each candidate is evaluated on a validation subset using the VLM, and its detection accuracy is recorded. The top-k high-performing instructions are retained, and the best-performing instruction among them is denoted as *Prompt-B*.

To encourage linguistic diversity while preserving semantic intent, a subset of the retained instructions is randomly selected as parent instructions and sent back to the LLM for transformation. For each selected parent instruction, the LLM performs either semantic rewriting, which preserves meaning while chang-

ing wording, or reasoning-expression modification, which changes the structural form of the instruction, each with probability 0.5. This process generates a set of refined child instructions.

The refined instructions are merged with the retained top-k parents to form the updated instruction pool while maintaining a fixed pool size of N. The generation-evaluation-refinement process is repeated for M rounds, gradually exploring diverse yet semantically related instruction formulations. At the end of optimization, the resulting pool provides a family of instruction variants for authenticity assessment, and the best-performing instruction *Prompt-B* is used as the default query in the following experiments.

3.3 Label-Conditioned Regeneration

Under binary authenticity supervision, VLLM training directly supervises the final authenticity prediction, but does not provide explicit supervision on the content of the generated text. To reuse the same binary supervision signal on the language side, without collecting human-written rationales, we introduce a label-conditioned regeneration mechanism.

Using the optimized instruction p_{best}, the VLM first produces a textual output and an authenticity prediction for each image x_j:

$$(r_j, \hat{y}_j) = f_{\text{VLM}}(x_j, p_{\text{best}}), \quad \hat{y}_j \in \{\text{Authentic}, \text{Generated}\}. \tag{6}$$

If the predicted label $\hat{y}_j$ matches the ground truth y_j, we retain the generated text r_j. Otherwise, we prepend a label directive d_j that explicitly specifies the target authenticity label, and regenerate the textual output under this constraint:

$$(r'_j, \hat{y}'_j) = f_{\text{VLM}}(x_j, p_{\text{best}}, d_j), \quad \hat{y}'_j = y_j. \tag{7}$$

The label directive is formulated as:

- `If the correct label is 'Generated':`
  `'This image is 'Generated' (AI-generated). Explain why this image is AI-generated, not authentic.'` + p_{best}
- `If the correct label is 'Authentic':`
  `'This image is 'Authentic' (not AI-generated). Explain why this image is authentic, not AI-generated.'` + p_{best}

This produces a regeneration dataset:

$$\mathcal{D}_{\text{regen}} = \{(x_j, r_j, y_j) \mid \hat{y}_j = y_j\} \cup \{(x_j, r'_j, y_j) \mid \hat{y}_j \neq y_j\}. \tag{8}$$

We then freeze the vision encoder and fine-tune the language module on $\mathcal{D}_{\text{regen}}$ using a standard autoregressive objective:

$$\mathcal{L}_{\text{text}} = -\sum_{(x,r,y)\in\mathcal{D}_{\text{regen}}} \log p_\theta(r \mid x, y, p_{\text{best}}), \tag{9}$$

where the visual encoder remains fixed and only language-side parameters are updated. In this way, binary authenticity labels provide scalable supervision

not only for the final prediction, but also for training the language-side textual outputs, without requiring human-annotated rationales.

Scope of Textual Outputs Because our supervision is limited to a binary authenticity label, the generated reasoning traces are not guaranteed to be faithful to image evidence, and hallucination may occur. We do not evaluate or claim faithfulness of these texts as verified explanations. In UC-VLM, they are used only as an auxiliary training signal to refine language-side behavior under the same authenticity supervision, supporting more consistent outputs for detection rather than ground-truth justification.

4 Experiments

Datasets and Metrics We evaluate **UC-VLM** on two challenging benchmarks, **GenImage** and **Chameleon**, which cover diverse real and synthetic image domains. Following the official protocols and prior works [34, 53, 62, 72], we report classification accuracy (ACC) as the primary evaluation metric.

- **GenImage** [72] is a large-scale benchmark for AIGI detection across diverse generative models. It contains over 2.68 million images, including 1.33 million real images from ImageNet [11] and 1.35 million generated images from eight models: BigGAN [8], GLIDE [41], VQDM, Stable Diffusion V1.4/V1.5 [49], ADM [12], Midjourney [5], and Wukong [6]. The dataset is organized into eight subsets by source generator, enabling per-model evaluation.
- **Chameleon** [62] is a high-resolution test-only benchmark with 26,000 images at resolutions ranging from 720P to 4K. Real images are collected from Unsplash [4], while AI-generated images are collected from platforms such as ArtStation [1], Civitai [2], and Liblib [3]. The synthetic samples are mainly produced by commercial systems such as Midjourney [5] and DALL·E 3 [47], or by custom LoRA [17]-based models built on Stable Diffusion [49]. Following the standard protocol, we report results under both ProGAN and Stable Diffusion V1.4 (SDV1.4) training settings.

Implementation Details We implement UC-VLM based on Qwen2.5-VL-7B [7], initializing both the vision encoder and the language model from the official checkpoint. All experiments are run on a single NVIDIA A100 GPU with mixed-precision (FP16) training. Instruction optimization is performed offline using GPT-4o: starting from a base template, we generate N=5 candidates and evolve them for M=10 rounds, retaining top-k ($k = 2$) by validation accuracy. The best instruction (*Prompt-B*) is selected from candidate instructions using sampled examples from the corresponding dataset, and is then fixed for both training and inference. For visual adaptation, we apply LoRA to the vision encoder (rank 16) and train for one epoch using SGD with learning rate 1×10^{-4} and batch size 32. For language-side adaptation, we fine-tune the language model with LoRA (rank 8) on $\mathcal{D}_{\text{regen}}$ for one epoch with learning rate 1×10^{-4}, while keeping the vision encoder fixed.

Table 2: Comparisons between the proposed UC-VLM and state-of-the-art methods on the GenImage testing set. The symbol †indicates our reproduction using the source code.

Method	Venue	Testing Subset								Avg
		Midjourney	SDV1.4	SDV1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Accuracy (%)
ResNet-50 [16]	CVPR'16	54.9	**99.9**	99.7	53.5	61.9	98.2	56.6	52.0	72.1
DeiT-S [56]	ICML'21	55.6	**99.9**	99.8	49.8	58.1	98.9	56.9	53.5	71.6
Swin-T [31]	ICCV'21	62.1	**99.9**	99.8	49.8	67.6	99.1	62.3	57.6	74.8
CNNSpot [57]	CVPR'20	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
Spec [68]	WIFS'19	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8
F3Net [45]	ECCV'20	50.1	**99.9**	**99.9**	49.9	50.0	**99.9**	49.9	49.9	68.7
GramNet [32]	CVPR'20	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9
DIRE [59]	ICCV'23	65.8	99.7	99.7	54.5	58.1	99.4	54.3	49.8	72.7
FatFormer† [30]	CVPR'24	93.1	97.0	97.2	82.0	95.0	95.8	88.8	49.9	87.4
NPR [53]	CVPR'24	81.0	98.2	97.9	76.9	89.8	96.9	84.1	84.2	88.6
DRCT [9]	ICML'24	91.5	95.0	94.4	79.4	89.2	94.7	90.0	81.7	89.5
VIB-Net [64]	CVPR'25	88.1	99.6	99.2	73.9	74.3	98.3	89.4	-	88.9
AIDE [62]	ICLR'25	79.4	99.7	99.8	78.5	91.8	98.7	80.3	66.9	86.9
DEUA [18]	ICCV'25	86.4	96.5	96.2	85.3	94.4	96.2	93.1	84.2	91.5
FIND [21]	AAAI'26	82.9	89.7	89.8	**87.4**	97.8	87.0	89.3	83.0	88.4
UC-VLM	-	**90.1**	99.9	99.8	87.0	**98.5**	96.9	**98.1**	**98.8**	**96.1**

Table 3: Comparisons between the proposed UC-VLM and state-of-the-art methods in terms of accuracy on the Chameleon dataset.

Training Set	Methods										
	CNNSpot [57]	FreDect [14]	Fusing [20]	GramNet [33]	LNP [28]	UniFD [43]	DIRE [59]	Patchcraft [70]	NPR [53]	AIDE [62]	UC-VLM
ProGAN	56.9	55.6	57.0	58.9	57.1	57.2	58.2	53.8	57.3	58.4	**69.6**
SDV1.4	60.1	56.9	57.1	61.0	55.6	55.6	59.7	56.3	58.1	62.6	**77.9**

4.1 Comparison with State-of-the-Arts

We compare UC-VLM with recent state-of-the-art methods on two benchmarks, **GenImage** and **Chameleon**. Table 2 reports the results on the GenImage test set. Conventional CNN-based and Transformer-based detectors [16,31,32,45,56, 57, 68] can achieve strong performance on specific generators, but their performance often degrades across different generative settings. More recent detectors such as DIRE [59], FatFormer [30], VIB-Net [64], NPR [53], DRCT [9], and DEUA [18] further improve accuracy by introducing reconstruction cues, frequency-aware priors, or sampling-related signals. As shown in Table 2, UC-VLM achieves an average accuracy of **96.1%** on GenImage, outperforming the strongest prior result (**91.5%** by DEUA) by **4.6%**. UC-VLM also delivers the best or among the best results on nearly all source generators, covering both diffusion-based and GAN-based models. These results indicate that the proposed unified binary-supervised framework can effectively adapt a VLLM to AIGI detection across diverse generative settings.

We further evaluate UC-VLM on the challenging Chameleon benchmark [62], which contains 26K high-resolution images spanning diverse content and generative sources. We compare against a broad set of detectors, including CNNSpot [57], FreDect [14], Fusing [20], LNP [28], UniFD [43], DIRE [59], Patchcraft [70], NPR [53], and AIDE [62]. As shown in Table 3, UC-VLM achieves the best performance under both training protocols, reaching **69.6%** under ProGAN training

Table 4: Ablation study of UC-VLM components. Baseline (Visual) and Baseline (LLM) denote fine-tuning only the vision encoder or the language module, respectively, while Baseline (Visual + LLM) denotes naive joint fine-tuning of both.

Setting	GenImage (%)	Chameleon (%)
Baseline Methods		
Baseline (w/o fine-tuning)	51.9	55.5
Baseline (Visual)	76.8	50.9
Baseline (LLM)	80.3	71.0
Baseline (Visual + LLM)	77.6	52.0
Component-wise Ablations		
Baseline (w/o fine-tuning) + Instruction Optimization	74.2	68.3
Baseline (Visual) + Visual Adaptation	91.7	55.4
Baseline (LLM) + Label-Conditioned Regeneration	82.4	65.1
UC-VLM	**96.1**	**77.9**

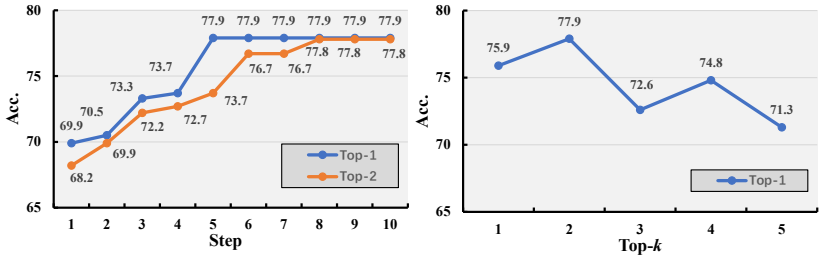

Fig. 3: Visualization of results on Chameleon under different step and Top-k. According to the results, we finally set the step to 10 and the k to 2.

and **77.9%** under SDV1.4 training. Compared with the strongest prior baseline, AIDE, UC-VLM improves accuracy by **11.2%** under ProGAN training and **15.3%** under SDV1.4 training. These results further support the effectiveness of UC-VLM on challenging high-quality synthetic images.

4.2 Ablation Studies

We conduct ablation experiments to analyze the contribution of each component in UC-VLM, as summarized in Table 4. Our goal is to examine how visual adaptation, instruction optimization, and label-conditioned regeneration each affect AIGI detection. Starting from the frozen backbone (first row), Qwen2.5-VL-7B shows limited discriminative ability on both GenImage and Chameleon, indicating that a generic VLLM is not directly optimized for subtle generative artifacts. Fine-tuning the visual encoder alone improves performance on GenImage but brings only limited gains on Chameleon, suggesting that artifact-sensitive visual features are useful but not sufficient for robust generalization. In contrast, adapting the language module yields larger improvements on both datasets, highlighting the importance of language-side refinement for authenticity prediction. Naively updating both the visual and language modules together does not produce consistent gains. This suggests that simply tuning both path-

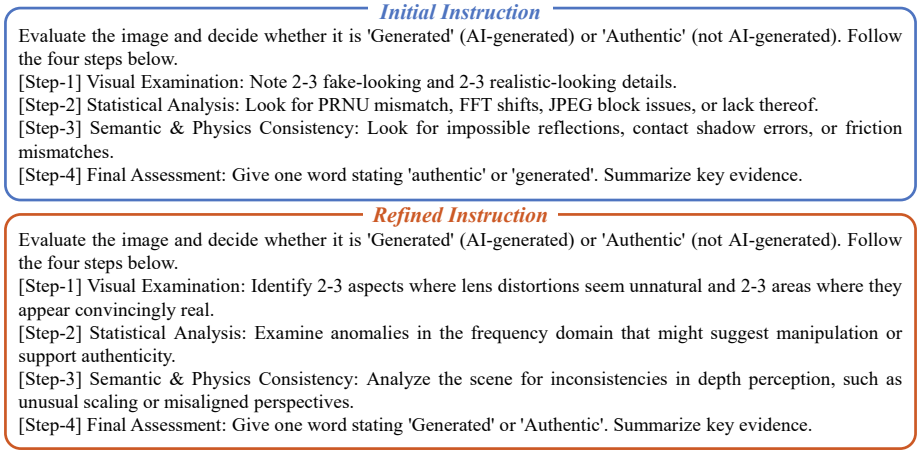

Fig. 4: Instruction comparison on the Chameleon dataset. After instruction optimization, UC-VLM's evolved prompt shifts from generic artifact descriptions to more specific visual cues, demonstrating stronger grounding in visual information.

ways under binary supervision is insufficient to coordinate visual and linguistic adaptation effectively.

We then evaluate each proposed component individually. Instruction optimization improves the frozen backbone, especially on Chameleon, confirming that VLLM predictions are sensitive to instruction formulation. Visual adaptation further strengthens the visual baseline by improving sensitivity to artifact-level cues. Label-conditioned regeneration improves the language baseline on GenImage, although its isolated effect is weaker on Chameleon. However, combining all components yields the full UC-VLM model, which achieves the best performance on both datasets (96.1% on GenImage and 77.9% on Chameleon). These results show that the three components are complementary, and that their integration leads to more robust AIGI detection than optimizing each component in isolation.

4.3 Discussion

Hyperparameters of Instruction Optimization Figure 3 analyzes the effect of two key hyperparameters in instruction optimization: the number of evolution rounds M and the Top-k selection size k. As M increases, detection performance improves and then gradually saturates, indicating that iterative instruction refinement helps identify more effective formulations for authenticity assessment.

This trend also reflects the sensitivity of VLLMs to prompt formulation. Although the backbone model already contains useful visual-textual knowledge, its detection performance depends strongly on how the task is queried. This motivates automatic instruction optimization instead of relying solely on manually designed prompts. A moderate Top-k value gives the best performance. When k is too small, the search tends to converge prematurely; when k is too large,

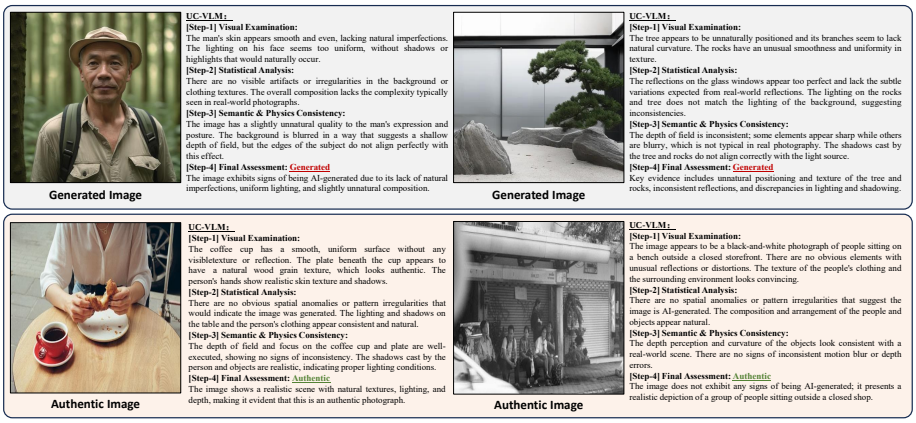

Fig. 5: Qualitative results of UC-VLM on Chameleon dataset. UC-VLM provides structured four-step explanations for both authentic and AI-generated images, often describing texture irregularities, lighting inconsistencies, depth errors, and semantic-physics mismatches. These examples illustrate its qualitative behavior.

the instruction pool contains more redundant and lower-quality variants, which weakens the refinement process. Overall, these results suggest that the proposed optimization process provides an effective balance between preserving strong candidate instructions and exploring diverse alternatives, enabling the model to discover more effective instruction variants for AIGI detection.

Instruction Variation for UC-VLM Figure 4 compares the best prompts obtained before and after Instruction Optimization. The earlier prompt tends to enumerate a scattered set of low-level cues, such as PRNU mismatch, FFT shifts, and JPEG block effects. While these cues are relevant to authenticity assessment, they appear broad and are not clearly prioritized. After instruction optimization, the optimized prompt becomes more structured and places greater emphasis on cues such as frequency patterns, lens distortion abnormalities, depth inconsistency, and perspective mismatch. This shift suggests that adapting the vision pathway changes the visual evidence the model relies on, which in turn affects which instruction formulations are most effective. Overall, these results indicate that Instruction Optimization moves the model from loosely collected artifact cues toward more task-oriented visual signals, providing a stronger basis for downstream detection and label-conditioned generation.

Qualitative Results Figure 5 shows qualitative examples that illustrate UC-VLM's behavior under the proposed unified learning framework. We emphasize that the generated text is not treated as a verified explanation. Instead, UC-VLM uses label-conditioned language-side refinement and optimized instructions as an auxiliary signal to stabilize predictions under binary supervision. Across examples, UC-VLM identifies visual irregularities commonly observed in synthesized content and maintains reliable predictions on authentic images. Compared with

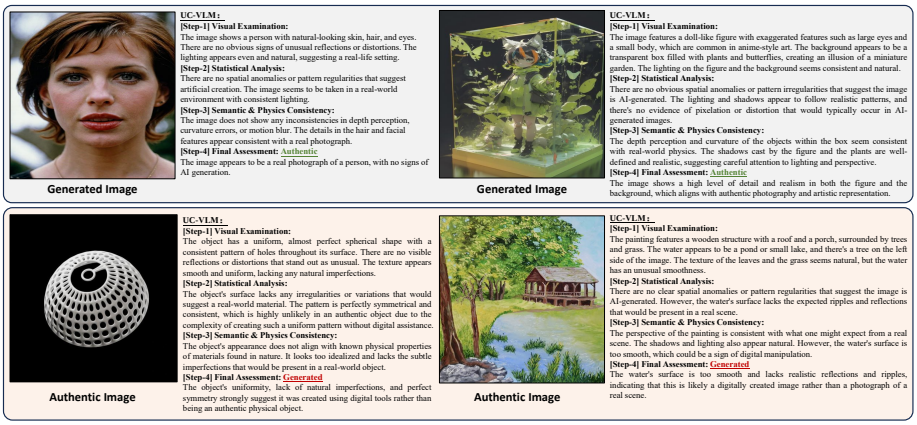

Fig. 6: Failure Cases of UC-VLM on Chameleon dataset. UC-VLM may misclassify highly realistic generated photos as authentic when artifacts are visually subtle, and may also mistake overly pristine or heavily post-processed authentic images. It is also uncertain on artistic or stylized imagery, where coherent generated artworks can be judged as authentic while authentic paintings can be judged as generated.

the backbone model, its outputs are more stable and are more often consistent with the target authenticity label across diverse cases.

Failure Cases In Figure 6, we show several typical failure cases of UC-VLM. First, highly realistic generated images with natural facial details, consistent lighting, and plausible geometry can be misclassified as authentic, since their synthesis artifacts are visually subtle. Second, the model shows clear uncertainty on artistic or non-photographic images: a generated digital artwork may be judged as authentic, while an authentic painting may be predicted as generated. These cases suggest that the detector has not fully learned an authenticity criterion independent of visual style and instead, it may still confuse photorealism or stylistic regularity with real/fake evidence.

5 Conclusion

We introduced **UC-VLM**, a unified learning framework for AI-generated image detection under binary authenticity supervision. UC-VLM integrates visual adaptation, robustness to instruction variation, and label-conditioned text generation within a unified multi-stage framework. Our key novelty lies in formulating these components as one binary-supervised learning problem, so that the same authenticity label is reused to constrain both the visual pathway and the language output without requiring human-annotated rationales or manual prompt engineering. This design strengthens sensitivity to low-level forensic cues, reduces sensitivity to prompt variation, and provides scalable supervision for textual outputs. Our results show that unified binary-supervised learning is an effective way to adapt VLLMs for AIGI detection.

Acknowledgment

This document is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 3 Grant (MOE-MOET32022-0001).

References

1. Artstation. https://www.artstation.com 10
2. Civitai. https://civitai.com 10
3. Liblib. https://www.liblib.art 10
4. Unsplash. https://unsplash.com 10
5. Midjourney. https://www.midjourney.com/home/ (2022) 10
6. Wukong. https://xihe.mindspore.cn/modelzoo/wukong (2022) 10
7. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 5, 10
8. Brock, A.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018) 10
9. Chen, B., Zeng, J., Yang, J., Yang, R.: Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In: Forty-first International Conference on Machine Learning (2024) 11
10. Cozzolino, D., Poggi, G., Nießner, M., Verdoliva, L.: Zero-shot detection of ai-generated images. In: European Conference on Computer Vision. pp. 54–72. Springer (2025) 2, 4
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009) 10
12. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 10
13. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024) 1
14. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: ICML. pp. 3247–3258. PMLR (2020) 4, 11
15. Gao, Y., Chang, D., Yu, B., Qin, H., Chen, L., Liang, K., Ma, Z.: Fakereasoning: Towards generalizable forgery detection and reasoning. arXiv preprint arXiv:2503.21210 (2025) 2, 4
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 4, 11
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022) 10
18. Huang, Y., Guo, H., Huang, J., Bai, B., Xiong, Q.: Diffusion epistemic uncertainty with asymmetric learning for diffusion-generated image detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17097–17107 (2025) 4, 11
19. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024) 2

20. Ju, Y., Jia, S., Ke, L., Xue, H., Nagano, K., Lyu, S.: Fusing global and local features for generalized ai-synthesized image detection. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 3465–3469. IEEE (2022) 11
21. Li, J., Feng, Y., Xie, C., Hu, J., Tan, L., Ji, J.: Find: A simple yet effective baseline for diffusion-generated image detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 6217–6225 (2026) 11
22. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: CVPR. pp. 5001–5010 (2020) 4
23. Li, Q., Wang, X.: Sponge tool attack: Stealthy denial-of-efficiency against tool-augmented agentic reasoning. arXiv preprint arXiv:2601.17566 (2026) 1
24. Li, Q., Wang, X.: Vimu: Benchmarking video metaphorical understanding. arXiv preprint arXiv:2605.14607 (2026) 1
25. Li, Q., Yin, B., Huang, W., Liu, R., Zou, B., Yu, R., Ye, J., Yu, W., Wang, X.: Vision-language-action safety: Threats, challenges, evaluations, and mechanisms. arXiv preprint arXiv:2604.23775 (2026) 1
26. Li, Q., Yu, R., Wang, X.: Vid-sme: Membership inference attacks against large video understanding models. Advances in Neural Information Processing Systems **38**, 111572–111596 (2026) 1
27. Li, S., Tan, L., Tan, R.T.: Bridging day and night: Target-class hallucination suppression in unpaired image translation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 6424–6432 (2026) 1
28. Liu, B., Yang, F., Bi, X., Xiao, B., Li, W., Gao, X.: Detecting generated images by real images. In: European Conference on Computer Vision. pp. 95–110. Springer (2022) 11
29. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023) 2
30. Liu, H., Tan, Z., Tan, C., Wei, Y., Wang, J., Zhao, Y.: 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) 4, 11
31. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021) 4, 11
32. Liu, Z., Qi, X., Torr, P.H.: Global texture enhancement for fake face detection in the wild. In: CVPR. pp. 8060–8069 (2020) 4, 11
33. Liu, Z., Qi, X., Torr, P.H.: Global texture enhancement for fake face detection in the wild. In: CVPR. pp. 8060–8069 (2020) 11
34. Luo, Y., Du, J., Yan, K., Ding, S.: Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17006–17015 (2024) 2, 4, 6, 10
35. Ma, Y., Jin, T., Zheng, X., Wang, Y., Li, H., Wu, Y., Jiang, G., Zhang, W., Ji, R.: Ompq: Orthogonal mixed precision quantization. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 9029–9037 (2023) 4
36. Ma, Y., Li, H., Zheng, X., Ling, F., Xiao, X., Wang, R., Wen, S., Chao, F., Ji, R.: Outlier-aware slicing for post-training quantization in vision transformer. In: Forty-first International Conference on Machine Learning (2024) 4
37. Ma, Y., Li, H., Zheng, X., Xiao, X., Wang, R., Wen, S., Pan, X., Chao, F., Ji, R.: Solving oscillation problem in post-training quantization through a theoretical

perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7950–7959 (2023) 4
38. Ma, Y., Zheng, X., Xu, J., Xu, X., Ling, F., Zheng, X., Kuang, H., Li, H., Wang, X., Xiao, X., et al.: Flow caching for autoregressive video generation. arXiv preprint arXiv:2602.10825 (2026) 1
39. Mayer, O., Stamm, M.C.: Accurate and efficient image forgery detection using lateral chromatic aberration. IEEE Transactions on information forensics and security **13**(7), 1762–1777 (2018) 4
40. McCloskey, S., Albright, M.: Detecting gan-generated imagery using saturation cues. In: ICIP. pp. 4584–4588. IEEE (2019) 4
41. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021) 10
42. O'brien, J.F., Farid, H.: Exposing photo manipulation with inconsistent reflections. ACM Trans. Graph. **31**(1), 4–1 (2012) 4
43. Ojha, U., Li, Y., Lee, Y.J.: 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) 11
44. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 1
45. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: ECCV. pp. 86–103. Springer (2020) 11
46. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021) 4
47. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 10
48. Ricker, J., Lukovnikov, D., Fischer, A.: 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) 2
49. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022) 10
50. Sarkar, A., Mai, H., Mahapatra, A., Lazebnik, S., Forsyth, D.A., Bhattad, A.: Shadows don't lie and lines can't bend! generative models don't know projective geometry... for now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 28140–28149 (2024) 2
51. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019) 1
52. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5052–5060 (2024) 2
53. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: 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) 2, 4, 6, 10, 11

54. Tan, L., Li, S., Kankanhalli, M., Tan, R.T.: Aggregating diverse cue experts for ai-generated image detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 9314–9322 (2026) 4
55. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems **37**, 84839–84865 (2024) 1
56. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021) 11
57. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: CVPR. pp. 8695–8704 (2020) 4, 11
58. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025) 2
59. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection. arXiv preprint arXiv:2303.09295 (2023) 2, 4, 11
60. Wen, S., Ye, J., Feng, P., Kang, H., Wen, Z., Chen, Y., Wu, J., Wu, W., He, C., Li, W.: Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. arXiv preprint arXiv:2503.14905 (2025) 2, 4
61. Wu, H., Zhou, J., Zhang, S.: Generalizable synthetic image detection via language-guided contrastive learning. arXiv preprint arXiv:2305.13800 (2023) 4
62. Yan, S., Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Xie, W.: A sanity check for ai-generated image detection. In: International Conference on Learning Representations (2025) 4, 10, 11
63. Yan, Z., Wang, J., Jin, P., Zhang, K.Y., Liu, C., Chen, S., Yao, T., Ding, S., Wu, B., Yuan, L.: Orthogonal subspace decomposition for generalizable ai-generated image detection. In: Forty-second International Conference on Machine Learning (2025) 4
64. Zhang, H., He, Q., Bi, X., Li, W., Liu, B., Xiao, B.: Towards universal ai-generated image detection by variational information bottleneck network. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23828–23837 (2025) 11
65. Zhang, X., Chen, Y.C.: Adaptive domain generalization via online disagreement minimization. IEEE Transactions on Image Processing **32**, 4247–4258 (2023) 4
66. Zhang, X., Tan, R.T.: Mamba as a bridge: Where vision foundation models meet vision language models for domain-generalized semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14527–14537 (2025) 2
67. Zhang, X., Xie, J., Yuan, Y., Mi, M.B., Tan, R.T.: Heap: unsupervised object discovery and localization with contrastive grouping. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 7323–7331 (2024) 4
68. Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in gan fake images. In: WIFS. pp. 1–6. IEEE (2019) 4, 11
69. Zhao, R., Li, W., Zhu, L., Zheng, Y., Lin, W.: Just noticeable difference modeling for deep visual features. In: International Conference on Machine Learning (2026) 4
70. Zhong, N., Xu, Y., Qian, Z., Zhang, X.: Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. CoRR (2023) 11
71. Zhou, Z., Luo, Y., Wu, Y., Sun, K., Ji, J., Yan, K., Ding, S., Sun, X., Wu, Y., Ji, R.: Aigi-holmes: Towards explainable and generalizable ai-generated image detection

via multimodal large language models. arXiv preprint arXiv:2507.02664 (2025) 2, 5

72. Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y.: Genimage: A million-scale benchmark for detecting ai-generated image. Advances in Neural Information Processing Systems **36** (2024) 10