SYNERGENDET: INTERNAL FLOW FORENSICS AND TIMESTEP ADAPTIVE ALIGNMENT FOR UNIFIED IMAGE GENERATION AND DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image generation and generated-image detection have both advanced rapidly, but they iterate independently for a long time, lacking a mechanism for mutual promotion and co-evolution. Although some recent work has attempted to use a unified architecture to achieve co-evolution of generation and detection, it has not fully leveraged the advantages of a unified architecture: detection remains a purely semantic judgment, while generation does not effectively exploit detection feedback. Therefore, we propose **SynerGenDet**: a unified framework that deeply couples generation and detection. On the detection side, we leverage the model's native generative capability for authenticity discrimination in a single unified forward pass, without requiring an auxiliary generative model or iterative denoising. Specifically, we use its one-step flow matching error as a generative forensic cue that complements semantic evidence. On the generation side, we introduce Timestep adaptive Alignment, which guides image generation with detector feedback at the right coarse-to-fine granularity. Experiments on multiple benchmarks show that SynerGenDet achieves state-of-the-art performance.

1 INTRODUCTION

The capabilities of generative models are improving at an unprecedented pace: from GANs (Jeong et al., 2022) and VAEs (Child, 2020) to diffusion models (Esser et al., 2024; Batifol et al., 2025), the realism and fidelity of generated content have been significantly enhanced, and recent unified generation-understanding models go further by synthesizing images through next-token prediction (Tian et al., 2024). With these models, "seeing is believing" will no longer hold true—people can generate almost any content they want, giving rise to new application forms such as AI short dramas and AI models. However, this also brings serious security risks: when generated content is indistinguishable from real content, illegal activities such as forgery and perjury may surge on a large scale, placing higher demands on AI content detection. Current detection methods (Ma et al., 2025a) mostly rely on visual semantic features to identify inconsistencies in content, performing reasonably well on known generators, but often failing to generalize well to new, unseen generators. Recent work (Xu et al., 2026a) brings generation and detection into a unified framework, creating opportunities for both tasks to benefit from shared representation.

Existing unified generative-understanding work (Ye et al., 2026) shows that generators can leverage understanding knowledge to generate higher-quality content that better meets user needs, while understanding can in turn interpret images more deeply through generative features. Therefore, unifying generation and detection is a natural and urgent direction in our view. However, in existing unified models (Zhang et al., 2026), generation and detection primarily share a common backbone, with limited interaction at the mechanistic level. The detection branch still operates largely as a discriminative classifier over semantic features and does not fully exploit the native generative capability of the shared model.

As illustrated in Figure 1, existing approaches typically establish only one-way knowledge transfer between generation and detection: a frozen generator provides forensic cues to a separate detector, or a frozen detector guides image generation as a fixed teacher. In contrast, we propose SynerGenDet, which enables the two capabilities to mutually reinforce each other within a unified model

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.

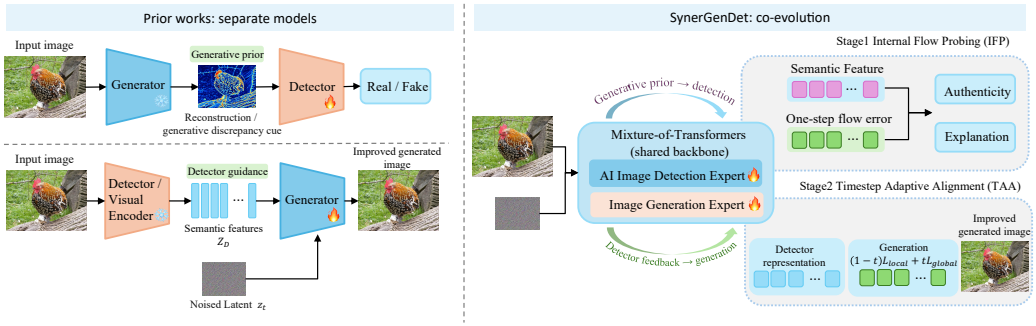


Figure 1: Comparison of existing one-way generation–detection paradigms and SynerGenDet. Unlike approaches that use a frozen generator or detector for one-way knowledge transfer, SynerGenDet couples both capabilities within a shared backbone, enabling two-stage mutual promotion.

through two lightweight mechanisms: (i) Internal Flow Probing (IFP), which re-injects the image to be detected as a noisy probe, using its single-step flow matching error to assist in authenticity judgment; (ii) Timestep Adaptive Alignment (TAA), which blends the detector-feature alignment target between global and local according to the timestep, so that generation receives feedback at the right granularity. Our experiments show improved detection, more balanced accuracy across unseen generators, and lower FID while retaining competitive compositional generation performance.

2 Related Work

Synthetic Image Detection. Synthetic image detection has evolved from relying on low-level artifacts (Wang et al., 2023) to fusing semantic features (Yang et al., 2026b), and then to explainable detection based on LLM (Xu et al., 2026b; Tan et al., 2026b). Early detectors exploit spectral, frequency-domain, reconstruction, or pixel-neighborhood artifacts (Wang et al., 2020; 2023; Tan et al., 2024a;b). More recent methods combine pretrained visual representations, such as CLIP and DINO, with forensic cues to improve generalization (Radford et al., 2021; Siméoni et al., 2025; Tan et al., 2025; Liu et al., 2024). Explainable detectors further generate human-interpretable rationales or employ retrieval and reinforcement learning to improve reasoning (Ji et al., 2026; Li et al., 2025b; Huang et al., 2025a; Tan et al., 2026a; Wen et al., 2025; Li et al., 2026). Nevertheless, these approaches remain primarily discriminative and rely largely on visual-semantic evidence, limiting their ability to exploit the native generative dynamics of a unified model.

Unified Understanding and Generation. Recent multimodal models unify understanding and generation by representing images as discrete or continuous visual tokens and jointly modeling visual and textual sequences (Team, 2024; Wang et al., 2024c; Chen et al., 2025; Wang et al., 2024a; Li et al., 2025a). Hybrid architectures, including JanusFlow and BAGEL, instead combine autoregressive language modeling with diffusion or flow-based image synthesis (Ma et al., 2025b; Deng et al., 2025). Because these models often retain separate semantic and generative visual spaces, subsequent work pursues shared or harmonized representations that preserve both high-level semantics and low-level visual details. (Qu et al., 2025; Jiao et al., 2025; Wu et al., 2025c; Wang et al., 2026).

Co-evolution of Generation and Detection. Generative responses can provide useful forensic evidence. DIRE, AEROBLADE, and LaRE^2 derive detection cues from image or latent reconstruction errors (Wang et al., 2023; Ricker et al., 2024; Luo et al., 2024), but rely on separate or fixed generative components. Conversely, REPA transfers semantic knowledge from a pretrained visual encoder to a diffusion transformer through representation alignment (Yu et al., 2024), while Dreaming in Flow enables mutual improvement between generation and visual understanding but does not address authenticity detection (Hao et al., 2026). UniGenDet more directly couples image generation and generated-image detection through joint training and detector-informed alignment (Zhang et al., 2026). However, it does not explicitly incorporate the model's internal flow-matching error as a forensic cue, and its detector-informed alignment is timestep-agnostic.

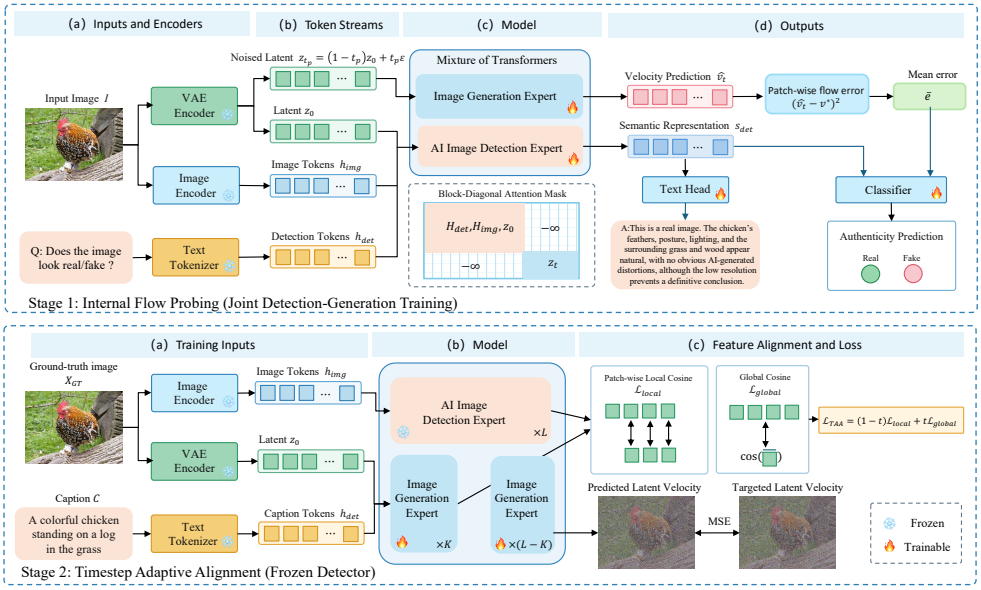


Figure 2: Overview of the two-stage SynerGenDet framework. Stage 1 uses Internal Flow Probing to combine semantic representations with one-step flow-matching errors for explainable authenticity detection. Stage 2 uses Timestep-Adaptive Alignment to guide generation with detector feedback at the appropriate coarse-to-fine granularity.

3 METHODOLOGY

We propose SynerGenDet, a unified generation–detection framework that promotes the co-evolution of both capabilities through a two-stage training strategy, as illustrated in Figure 2. Stage 1 jointly trains detection and generation, using an internal flow-error cue to improve authenticity classification. Stage 2 freezes the detection expert and uses detector representations to guide generation.

3.1 PRELIMINARIES

Let $z_0 \in \mathbb{R}^{N\times C}$ denote the patchified VAE latent of an image. Rectified flow defines a linear path between z_0 and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$:

$$z_t = (1-t)z_0 + t\epsilon, \qquad v^* = \frac{\mathrm{d}z_t}{\mathrm{d}t} = \epsilon - z_0, \qquad t \sim \mathcal{U}(0,1). \tag{1}$$

Given (z_t, t), the unified decoder predicts the velocity $\hat{v}_\theta(z_t, t) = W_v h_\theta(z_t, t)$ and is trained using

$$\mathcal{L}_{\text{fm}} = \mathbb{E}_{z_0,\epsilon,t}\left[\frac{1}{NC}\left\|\hat{v}_\theta(z_t,t) - v^*\right\|_F^2\right]. \tag{2}$$

We omit text conditioning for clarity. We use $\mathrm{sg}(\cdot)$ to denote the stop-gradient operator and $\sigma(\cdot)$ to denote the logistic sigmoid function.

3.2 STAGE 1: INTERNAL FLOW PROBING

Motivation. A unified model's generative pathway provides a complementary forensic prior beyond semantic representations. Reconstruction-based detectors, including DIRE, AEROBLADE, and LaRE^2, show that generative discrepancies contain useful forensic evidence, but typically require a separate or frozen generator and additional reconstruction or denoising. Internal Flow Probing (IFP) instead re-injects the input image as a noised latent probe and uses the unified model's one-step velocity-prediction error as a forensic signal complementary to semantic evidence.

During this stage, we jointly fine-tune the unified model on detection and generation data, with detection as the primary task. Keeping the generation pathway trainable allows its flow field to

remain meaningful while the model adapts to generated-image detection, providing complementary generative cues for robust authenticity discrimination. For each detection sample, we consider a tuple (I, Q, y, A), where I is the input image, Q is the detection instruction, $y \in \{0, 1\}$ is the binary authenticity label, and $A = (a_1, \ldots, a_T)$ is the corresponding natural-language explanation.

Dual Visual Representations. The input image I is encoded in two complementary visual spaces. An understanding encoder E_{und} (SigLIP2) produces semantic visual tokens, while a generation encoder E_{gen} (VAE) produces the clean generation latent:

$$h^0_{\text{und}} = E_{\text{und}}(I), \qquad z_0 = E_{\text{gen}}(I). \tag{3}$$

The instruction Q is embedded as text tokens h^0_{txt}. During training, the ground-truth explanation tokens are provided under teacher forcing; during inference, the explanation is decoded autoregressively. To probe the model's generative response, we perturb the clean latent z_0 at a fixed probe timestep t_p:

$$z_{t_p} = (1 - t_p)z_0 + t_p\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I). \tag{4}$$

Isolated Detection and Probe Streams. We process the input through two parallel token streams within a single packed decoder forward pass. The detection stream contains the text tokens, the visual-understanding tokens, and the clean generation tokens:

$$\mathcal{S}^0_{\text{det}} = \left[h^0_{\text{txt}}; h^0_{\text{und}}; z^0_0\right]. \tag{5}$$

These tokens jointly provide textual, semantic, and clean generative context for authenticity reasoning. In parallel, the noised latent z_{t_p} constitutes an isolated probe stream, $\mathcal{S}^0_{\text{probe}} = z^0_{t_p}$.

The probe tokens attend to one another, allowing the generation pathway to model spatial dependencies within the noised latent, but they do not attend to the detection stream. Conversely, tokens in the detection stream do not attend to the probe stream. We implement this separation using a block-diagonal attention mask:

$$M = \begin{bmatrix} M_{\text{det}} & -\infty \\ -\infty & M_{\text{probe}} \end{bmatrix}, \tag{6}$$

where M_{det} permits interactions within the detection stream and M_{probe} permits interactions among the noised probe tokens.

Under the Mixture-of-Transformers architecture, understanding and generation tokens are routed through their respective expert parameters. At layer ℓ, the detection and probe streams are packed into a complete token sequence and processed in a single forward pass:

$$\mathcal{S}^\ell = \left[\mathcal{S}^\ell_{\text{det}}; \mathcal{S}^\ell_{\text{probe}}\right], \qquad \mathcal{S}^{\ell+1} = D^\ell_\theta\left(\mathcal{S}^\ell, \tau; M\right). \tag{7}$$

where M is the block-diagonal attention mask defined in Eq. 6. The timestep assignment τ associates the noised probe tokens with t_p and the clean generation tokens with $t = 0$; text and understanding tokens do not receive a flow timestep. Within each layer, tokens are routed to the understanding or generation expert according to their token types.

This isolation prevents the velocity predictor from directly accessing the corresponding clean latent z_0, from which the flow-matching target could otherwise be trivially recovered. We use the final-layer hidden state that predicts the first output token as the detection representation $h_{\text{det}} = H^L_{\text{ans},1}$.

Internal Flow-error Readout. Let H^L_{probe} denote the final-layer hidden states of the probe tokens. The velocity head defined in (Eq. 1) produces a one-step prediction, which is compared with the stop-gradient flow-matching target:

$$\hat{v} = v_\theta(H^L_{\text{probe}}, t_p), \qquad v^* = \text{sg}(\epsilon - z_0). \tag{8}$$

to obtain the patch-wise flow-matching errors:

$$e_i = \left(\hat{v}_i - v^*_i\right)^2, \qquad i = 1, \ldots, N. \tag{9}$$

Here, $e_i \in \mathbb{R}^C$ is the channel-wise squared velocity error of the i-th latent patch, and $\text{sg}(\cdot)$ denotes the stop-gradient operator. We average the patch-wise errors to obtain a fixed-dimensional forensic feature, summarizing generative discrepancies across all latent spatial locations:

$$\bar{e} = \frac{1}{N}\sum_{i=1}^{N} e_i \in \mathbb{R}^C. \tag{10}$$

The isolated probe stream therefore contributes to detection only through its pooled flow-matching error, rather than through direct token-level attention.

Authenticity Prediction and Explanation. We concatenate the semantic detection representation h_{det} with the pooled flow-error feature $\bar{e}$ and pass them through an MLP head F_{det}:

$$\hat{y} = \sigma\left(F_{\text{det}}\left([h_{\text{det}}; \text{sg}(\bar{e})]\right)\right), \tag{11}$$

where $\sigma(\cdot)$ is the logistic sigmoid function. The explanation A is decoded autoregressively from the detection stream. Although the two token streams remain isolated inside the decoder, their complementary semantic and generative evidence is fused by the final authenticity-prediction head.

Training Objective. The Stage 1 objective combines authenticity classification, explanation generation, and standard flow-matching training:

$$\mathcal{L}_{\text{IFP}} = \lambda_{\text{det}}\mathcal{L}_{\text{det}} + \lambda_{\text{exp}}\mathcal{L}_{\text{exp}} + \lambda_{\text{fm}}\mathcal{L}_{\text{fm}}, \tag{12}$$

where

$$\mathcal{L}_{\text{det}} = \text{BCE}(\hat{y}, y), \qquad \mathcal{L}_{\text{exp}} = -\sum_{j=1}^{T} \log p_\theta(a_j \mid a_{<j}), \tag{13}$$

and $\mathcal{L}_{\text{fm}}$ is the standard flow-matching loss defined in Eq. 2. It is applied to the generation tokens from the generation-training samples, rather than to the isolated detection probes.

We stop the classification gradient at $\bar{e}$, so the detection objective does not directly update the probe stream. The shared generation pathway is instead maintained by the flow-matching loss computed on the generation-training samples, and its prediction error is used as a fixed generative cue for authenticity classification.

IFP requires neither an auxiliary generative model nor iterative denoising. The detection and probe streams are processed in one packed decoder forward pass, after which the semantic detection feature and the one-step flow-error feature are fused by a lightweight prediction head for authenticity classification. Natural-language explanations require subsequent autoregressive decoding.

3.3 Stage 2: Timestep Adaptive Alignment

Motivation. IFP trains the detector to encode features useful for authenticity discrimination, which can subsequently provide supervision for the generator. UniGenDet takes a step in this direction through detector-informed generative alignment, but applies a timestep-agnostic alignment objective throughout the generation trajectory. This design overlooks the coarse-to-fine nature of rectified-flow generation: high-noise timesteps primarily determine global semantic structure, whereas low-noise timesteps progressively refine local appearance and texture. Consequently, patch-level alignment may be unreliable when the latent is dominated by noise, while purely global alignment is too coarse to guide fine-grained refinement. We therefore introduce Timestep-Adaptive Alignment (TAA), which smoothly shifts the detector supervision from global structure at high-noise timesteps to local details at low-noise timesteps.

Let f_D denote the detector obtained from Stage 1, which is frozen throughout Stage 2, and let g_θ denote the trainable generator. Given a ground-truth image-caption pair (x_{GT}, c), we encode the image as $z_0 = E_g(x_{\text{GT}})$ and construct the noised latent z_t following (Eq. 1). For the b-th training sample, we extract the generator's intermediate patch features at layer K, where $1 \le K \le \ell$:

$$z_b^G = g_\theta^{(K)}(z_{t_b}, t_b, c_b) \in \mathbb{R}^{N_G \times C_G}. \tag{14}$$

We also extract patch-level representations from an intermediate layer of the frozen detector:

$$\widetilde{z}_b^D = f_D^{(K_D)}(x_{\text{GT},b}) \in \mathbb{R}^{N_D \times C_D}, \qquad z_b^D = \text{Resize}(\widetilde{z}_b^D) \in \mathbb{R}^{N_G \times C_D}, \tag{15}$$

where $\text{Resize}(\cdot)$ bilinearly resamples the detector features onto the generator's spatial patch grid. Because f_D is frozen, z_b^D serves as a fixed alignment target. A lightweight trainable projection h_ϕ maps the generator features into the detector feature space:

$$\hat{z}_b^G = h_\phi(z_b^G) \in \mathbb{R}^{N_G \times C_D}. \tag{16}$$

Using $\cos(u, v) = \langle u, v\rangle/(\|u\|_2\|v\|_2)$, we define the local alignment loss by matching spatially corresponding patches:

$$\mathcal{L}_{\text{local}}^{(b)} = \frac{1}{N_G}\sum_{i=1}^{N_G}\left[1 - \cos\left(\hat{z}_{b,i}^{G}, z_{b,i}^{D}\right)\right]. \tag{17}$$

For global alignment, we first average-pool the generator and detector features and then align their image-level representations:

$$\bar{z}_b^G = \frac{1}{N_G}\sum_{i=1}^{N_G}\hat{z}_{b,i}^{G}, \qquad \bar{z}_b^D = \frac{1}{N_G}\sum_{i=1}^{N_G} z_{b,i}^{D}, \qquad \mathcal{L}_{\text{global}}^{(b)} = 1 - \cos(\bar{z}_b^G, \bar{z}_b^D). \tag{18}$$

Rather than assigning a single weight based on the batch-average timestep, TAA adapts the alignment granularity independently for each sample:

$$\mathcal{L}_{\text{TAA}} = \frac{1}{B}\sum_{b=1}^{B}\left[(1-t_b)\mathcal{L}_{\text{local}}^{(b)} + t_b\mathcal{L}_{\text{global}}^{(b)}\right]. \tag{19}$$

Because $t_b = 1$ corresponds to pure noise and $t_b = 0$ to the clean latent, this weighting emphasizes global alignment at high-noise timesteps and gradually shifts toward patch-level alignment as the generation trajectory approaches the clean image.

Loss Function Design. The Stage 2 training objective combines the standard flow-matching loss with the proposed timestep-adaptive alignment objective:

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{fm}} + \lambda_{\text{TAA}}\mathcal{L}_{\text{TAA}}. \tag{20}$$

Along the rectified-flow path, high-noise states primarily encode coarse semantic structure, whereas low-noise states contain increasingly detailed spatial and textural information. TAA therefore matches the granularity of detector supervision to the generation state: global representations guide early structural formation, while spatially aligned patch features guide later detail refinement.

4 Experiments

4.1 Experimental Setup

Datasets. We use FakeClue (Wen et al., 2026), which provides binary authenticity labels and natural-language explanations, for detection training. Following BAGEL (Deng et al., 2025), we train the generation pathway on approximately 80K image–text pairs from the high-aesthetic subset of LAION (Schuhmann et al., 2022). We evaluate detection in domain on FakeClue, cross-dataset generalization on DMImage (Corvi et al., 2023) without fine-tuning, and zero-shot generalization to visual autoregressive generators on ARForensics (Zhang et al., 2025). Generation is evaluated using FID on 5,000 held-out LAION prompts and GenEval (Ghosh et al., 2023), which probes text-to-image alignment over 500+ prompts across six dimensions: single-object (SO), two-object (TO), counting (CT), colors (CL), position (POS), and attribute binding (ATTR).

Baselines. We compare SynerGenDet with general-purpose multimodal models, specialized synthetic-image detectors, and explainable detectors. The complete baseline descriptions and references are provided in Appendix A.1.

4.2 Explainable Detection and Generalization

In-domain Explainable Detection. We first evaluate SynerGenDet on the FakeClue test set. As shown in Table 1, we report authenticity classification using accuracy (Acc) and F1 score, and evaluate the generated explanations using ROUGE-L and CSS. ROUGE-L measures lexical overlap with the reference explanations, whereas CSS evaluates the semantic consistency between an explanation and its corresponding image.

Under the matched Stage-1 protocol, adding IFP improves UniGenDet's Acc from 98.0% to 99.0% and F1 from 97.7% to 98.8%. This controlled comparison supports the effectiveness of IFP for

Table 1: Results on FakeClue. * denotes methods trained on FakeClue.

Type	Method	FakeClue			
		Acc↑	F1↑	ROUGE-L↑	CSS↑
I	DeepSeek-VL2	47.5	54.1	17.2	50.5
	InternVL2-40B	50.7	46.3	17.6	55.2
	Qwen2-VL-72B	57.8	56.5	17.5	54.4
	Qwen3-VL-plus	58.9	58.4	21.5	46.2
	Qwen3.5-397B-A17B	67.0	66.9	22.9	49.8
	BAGEL	40.5	34.1	23.9	46.2
II	CNNSpot	43.1	9.8	–	–
	FatFormer	54.5	45.1	–	–
	UnivFD	63.1	46.8	–	–
	NPR*	90.2	91.6	–	–
	Fake-HR1*	86.9	87.1	–	–
	IFM-AIGCSpotter-3B*	97.1	97.7	–	–
	DF-LLaVA*	94.5	96.1	56.8	**89.7**
	FakeVLM*	98.6	98.1	32.2	59.5
	UniGenDet*	98.0	97.7	56.3	79.8
	SynerGenDet*	**99.0**	**98.8**	**57.1**	78.5

Table 2: Cross-dataset evaluation on DMImage using the original model weights.

Method	Real		Fake		Overall	
	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
CNNSpot	87.8	88.4	28.4	44.2	40.6	43.3
Gram-Net	62.8	54.1	78.8	88.1	67.4	79.4
Fusing	87.7	86.1	15.5	27.2	40.4	36.5
LNP	63.1	67.4	56.9	72.5	58.2	68.3
UnivFD	89.4	88.3	44.9	61.2	53.9	60.7
DF-LLaVA	74.8	85.6	96.0	98.0	91.0	94.2
AntiFakePrompt	91.3	92.5	89.3	91.2	90.6	91.2
SIDA	92.9	93.1	90.7	91.0	91.8	92.4
FakeVLM	98.2	99.1	89.7	94.6	94.0	94.3
UniGenDet	**99.0**	**99.5**	97.2	98.6	98.6	99.1
SynerGenDet	97.5	98.7	**99.9**	**99.6**	**99.3**	**99.2**

Table 3: Zero-shot generalization on ARForensics (Acc). I: trained on the LlamaGen dataset; II: zero-shot evaluation.

Type	Method	LlamaGen Acc.	VAR Acc.	Infinity Acc.	Janus-Pro Acc.	RAR Acc.	Switti Acc.	OM2 Acc.	Mean Acc.
I	CNNSpot	99.9	50.3	50.9	95.7	50.8	56.6	50.1	64.9
	FreDetect	99.8	52.9	50.2	88.9	52.5	50.0	57.1	64.5
	Gram-Net	99.6	55.0	52.4	74.5	50.0	57.7	50.1	62.8
	LNP	99.5	49.6	49.8	99.5	49.7	70.3	49.6	66.9
	UnivFD	89.9	80.5	71.7	84.3	88.3	76.0	66.2	79.6
	NPR	**100.0**	56.9	88.5	93.7	52.3	52.0	63.0	72.3
	D^3QE	97.2	85.3	62.9	92.3	91.7	75.3	70.1	82.1
II	FakeVLM	98.1	97.7	99.4	**99.9**	**99.9**	94.6	90.4	97.1
	UniGenDet	89.4	**99.7**	**99.9**	99.7	99.5	98.8	**99.5**	98.1
	Ours	97.5	97.7	98.2	98.3	98.7	**98.9**	97.9	**98.2**

authenticity detection. Among the evaluated general-purpose LMMs, the strongest accuracy is obtained by Qwen3.5-397b-A17B at 67.0%, substantially below our result. SynerGenDet also exceeds NPR and Fake-HR1 in accuracy by 8.8 and 12.1 percentage points, respectively. Compared with FakeVLM, our method improves Acc from 98.6% to 99.0% and F1 from 98.1% to 98.8%, while increasing ROUGE-L from 32.2% to 57.1%.

Our method obtains a CSS score of 78.5%, which is comparable to the 79.8% achieved by UniGenDet. These results indicate that the improved detection performance and reference-level explanation quality are obtained while largely preserving image–explanation semantic consistency. The results support the effectiveness of the one-step flow-matching error as a complementary forensic signal.

Cross-dataset Generalization. We also present the generalization experiment results of DFLLaVA on DMimage (Corvi et al., 2023), following (Zhang et al., 2026), without additional fine-tuning on DMImage. As reported in Table 2, SynerGenDet achieves an overall accuracy of 99.3% and an overall F1 score of 99.2%. These scores exceed those of UniGenDet by 0.7 and 0.1 percentage points, respectively. SynerGenDet also achieves higher accuracy and F1 than SIDA(91.8% and 92.4%) and FakeVLM (94.0% and 94.3%). These results suggest that SynerGenDet generalizes effectively beyond the detection training dataset.

When evaluated by class, SynerGenDet obtains 97.5% Acc on real images and 99.9% Acc on generated images, demonstrating strong performance on both classes rather than relying on a bias toward generated samples. The corresponding real- and generated-image F1 scores are 98.7% and 99.6%,

Table 4: Evaluation of text-to-image generation on GenEval benchmark. 'Gen. Only' stands for an image generation model, and 'Unified' denotes a model that has both understanding and generation capabilities. † refer to the methods using LLM rewriter.

Type	Model	SO	TO	CT	CL	POS	ATTR
Gen. Only	PixArt-Σ	0.98	0.50	0.44	0.80	0.08	0.07
	SDv2.1	0.98	0.51	0.44	0.85	0.07	0.17
	DALL-E 2	0.94	0.66	0.49	0.77	0.10	0.19
	SDXL	0.98	0.74	0.39	0.85	0.15	0.23
	DALL-E 3	0.96	0.87	0.47	0.83	0.43	0.45
	SD3-Medium	**0.99**	0.94	0.72	0.89	0.33	0.60
	FLUX.1-dev†	0.98	0.93	0.75	0.93	0.68	0.65
Unified	TokenFlow-XL	0.95	0.60	0.41	0.81	0.16	0.24
	Janus	0.97	0.68	0.30	0.84	0.46	0.42
	Emu3-Gen†	0.99	0.81	0.42	0.80	0.49	0.45
	Show-o	0.98	0.80	0.66	0.84	0.31	0.50
	OmniGen2	**1.00**	0.95	0.64	0.88	0.55	0.76
	Janus-Pro-7B	0.99	0.89	0.59	0.90	**0.79**	0.66
	BAGEL†	0.98	**0.97**	**0.86**	0.93	0.74	**0.77**
	UniGenDet†	0.99	0.95	0.80	0.94	0.74	0.75
	Ours†	**1.00**	0.94	0.82	**0.95**	0.78	**0.77**

respectively. These results demonstrate that the detector transfers effectively across datasets while retaining its natural-language explanation capability.

Generalization to Unseen Generators. Finally, we evaluate zero-shot generalization on ARForensics (Zhang et al., 2025), which contains images synthesized by seven visual autoregressive generators that are not observed during training. As shown in Table 3, SynerGenDet achieves a mean accuracy of 98.2%, compared with 98.1% for UniGenDet and 97.1% for FakeVLM. It also substantially outperforms specialized detectors trained on LlamaGen, including D^3QE, which obtains a mean accuracy of 82.1%.

Beyond the mean score, SynerGenDet improves the minimum accuracy across the seven evaluated generators from UniGenDet's 89.4% to 97.5%, an increase of 8.1 percentage points. Although UniGenDet retains higher accuracy on several generators, SynerGenDet exhibits a more balanced performance profile and improves the weakest-generator result. Across the seven unseen generators, the accuracy of SynerGenDet ranges from 97.5% to 98.9%, indicating stable performance across different autoregressive generation architectures. These results suggest that probing the model's internal flow dynamics provides a transferable forensic signal beyond generator-specific artifact patterns.

Table 5: Effect of the probe timestep t_p on FakeClue detection performance.

Probe timestep	Acc↑	F1↑
$t_p = 0.1$	98.2	97.7
$t_p = 0.3$	98.7	98.5
$t_p = 0.5$	**99.0**	**98.8**
$t_p = 0.8$	98.2	98.3

Table 6: Image quality comparison using FID. All alignment variants are built upon BAGEL + IFP. The fixed combination assigns equal weights to local and global alignment.

Variant	BAGEL	BAGEL + IFP	+ Local	+ Global	+ Local & Global (fixed)	+ TAA
FID↓	22.5	18.0	17.4	17.6	17.3	**16.9**

4.3 Generation Evaluation

Text-Image Alignment Assessment. We evaluate text-to-image generation performance on GenEval. As shown in Table 4, SynerGenDet achieves scores of 1.00, 0.94, 0.82, 0.95, 0.78, and 0.77

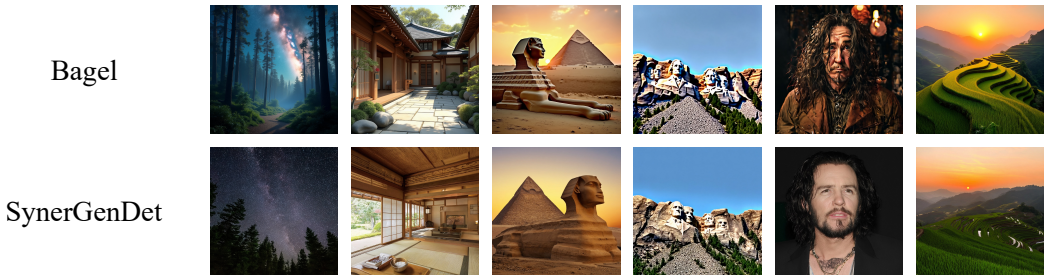


Figure 3: Qualitative comparison of BAGEL and SynerGenDet under the same prompts. SynerGenDet produces coherent and realistic images across diverse scenes.

on SO, TO, CT, CL, POS, and ATTR, respectively. Its overall score of 0.88 exceeds UniGenDet's 0.86 and remains comparable to that of the BAGEL backbone. Among the unified models evaluated, SynerGenDet achieves the highest CL score and ties for the highest scores on SO and ATTR. Compared with UniGenDet, it improves POS from 0.74 to 0.78 and ATTR from 0.75 to 0.77. Relative to BAGEL, it improves SO, CL, and POS, matches ATTR, and obtains lower scores on TO and CT. Together with the detection results, these findings indicate that SynerGenDet improves authenticity detection while maintaining competitive compositional generation performance.

4.4 Ablation Studies

To examine the contribution of each component and the sensitivity to key hyperparameters, we conduct ablation studies on both detection and generation.

Effect of the Probe Timestep. We investigate IFP's sensitivity to the probe timestep t_p, which controls the noise level applied to the clean generation latent. As shown in Table 5, Acc/F1 improve from 98.2%/97.7% at $t_p = 0.1$ to 98.7%/98.5% at $t_p = 0.3$, peaking at 99.0%/98.8% at $t_p = 0.5$. Increasing t_p to 0.8 reduces Acc/F1 to 98.2%/98.3%. These results suggest that an intermediate noise level provides the most discriminative flow-error cues among the tested settings. A possible explanation is that low noise provides insufficient perturbation, whereas excessive noise obscures image-specific information needed for authenticity discrimination. We therefore set $t_p = 0.5$ in all other experiments.

Effect of Detector-informed Alignment. We evaluate the distributional quality of generated images using Fréchet Inception Distance (FID). As shown in Table 6, joint Stage-1 training with IFP reduces FID from 22.5 for BAGEL to 18.0. Starting from this BAGEL + IFP checkpoint, local-only and global-only alignment further reduce FID to 17.4 and 17.6, respectively. Combining local and global alignment with equal, timestep-independent weights (0.5 each) achieves a FID of 17.3, whereas TAA achieves the lowest FID of 16.9. The additional 0.4-point reduction over the fixed-weight combination supports the benefit of adapting the relative contributions of local and global supervision along the flow trajectory. Overall, TAA improves FID by 1.1 points over BAGEL + IFP, while the full framework reduces FID by 5.6 points relative to BAGEL. Together with the GenEval results in Table 4, these findings indicate improved distributional image quality while maintaining strong compositional generation performance.

5 Conclusion

We presented SynerGenDet, a unified framework in which image generation and generated-image detection mutually reinforce each other. For detection, Internal Flow Probing (IFP) uses the model's one-step flow-matching error as a forensic cue complementary to semantic features, without an external generator or iterative denoising. For generation, Timestep-Adaptive Alignment (TAA) transfers authenticity-aware detector representations at a granularity matched to the coarse-to-fine flow trajectory. SynerGenDet achieves strong explainable detection and cross-generator generalization, with at least 97.5% accuracy on each of seven unseen generators, while improving FID and preserving compositional generation performance. These results demonstrate the potential of generative dynamics and forensic representations to benefit each other within a unified model.

AI USE STATEMENT

In this work, we used generative AI tools to assist with refining and editing our initial code implementations and with improving the grammar, clarity, and readability of our manuscript. We developed the research ideas, hypotheses, methodology, experimental design, and initial code implementations independently.

We manually reviewed all AI-assisted code to verify its logic and consistency with our intended methodology. Multiple authors participated in this review. We also reviewed and revised all AI-assisted text and verified the technical statements, experimental results, citations, and conclusions in the final manuscript. We did not use generative AI tools to generate datasets, formulate scientific claims, interpret experimental results, or create the core scientific content of this work. We take responsibility for the final content of this work, including all text and code produced or modified with the aid of generative AI.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

S Batifol, A Blattmann, F Boesel, S Consul, C Diagne, T Dockhorn, J English, Z English, P Esser, S Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. arxiv 2025. *arXiv preprint arXiv:2506.15742*, 2025.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Xiuli Bi, Bo Liu, Fan Yang, Bin Xiao, Weisheng Li, Gao Huang, and Pamela C Cosman. Detecting generated images by real images only. *arXiv preprint arXiv:2311.00962*, 2023.

You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419*, 2023.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-Σ: Weak-to-strong training of diffusion transformer for 4K text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024a.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

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

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *URL https://arxiv. org/abs/2505.14683*, 8, 2025.

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

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Dhruba Ghosh, Hanna Hajishirzi, and L Schmidt. An object-focused framework for evaluating text-to-image alignment. *NeurIPS*, 2023.

Ke Hao, Yuanzhi Liang, Tingxi Chen, Rui Li, Haibin Huang, Chi Zhang, Yun Gu, and Xuelong Li. Dreaming in flow: Generative grounding feedback for self-evolving unified multimodal models. *arXiv preprint arXiv:2609.08282*, 2026.

Tai-Ming Huang, Wei-Tung Lin, Kai-Lung Hua, Wen-Huang Cheng, Junichi Yamagishi, and Jun-Cheng Chen. Thinkfake: Reasoning in multimodal large language models for ai-generated image detection. *arXiv preprint arXiv:2509.19841*, 2025a.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28831–28841. IEEE, 2025b.

Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi. Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1060–1068, 2022.

Yikun Ji, Yan Hong, Qi Fan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, Jianfu Zhang, et al. Fakexplain: Ai-generated image detection via human-aligned grounded reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 68549–68570, 2026.

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 *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10482–10486. IEEE, 2026.

Yang Jiao, Haibo Qiu, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. Unitoken: Harmonizing multimodal understanding and generation through unified visual encoding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3600–3610. IEEE, 2025.

Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. 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.

Hao Li, Changyao Tian, Jie Shao, Xizhou Zhu, Zhaokai Wang, Jinguo Zhu, Wenhan Dou, Xiaogang Wang, Hongsheng Li, Lewei Lu, et al. Synergen-vl: Towards synergistic image understanding and generation with vision experts and token folding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29767–29779. IEEE, 2025a.

Tianxiao Li, Zhenglin Huang, Haiquan Wen, Yiwei He, Shuchang Lyu, Baoyuan Wu, and Guangliang Cheng. Raidx: A retrieval-augmented generation and grpo reinforcement learning framework for explainable deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11746–11755, 2025b.

Tianxiao Li, Zhenglin Huang, Haiquan Wen, Yiwei He, Xinze Li, Bingyu Zhu, Wuhui Duan, Congang Chen, Zeyu Fu, Yi Dong, et al. Omni-fake: benchmarking unified multimodal social media deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 30299–30311, 2026.

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

Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8057–8066. IEEE, 2020.

Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17006–17015. IEEE, 2024.

Long Ma, Zhiyuan Yan, Qinglang Guo, Yong Liao, Haiyang Yu, and Pengyuan Zhou. Detecting ai-generated video via frame consistency. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025a.

Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai Yu, et al. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7739–7751. IEEE, 2025b.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2545–2555. IEEE, 2025.

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

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

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

Zhuokang Shen, Kaisen Zhang, Bohan Jia, Heming Jia, Yuan Fang, Zhou Yu, and Shaohui Lin. DF-LLaVA: Unlocking MLLMs for synthetic image detection via knowledge injection and conflict-driven self-reflection. *arXiv preprint arXiv:2509.14957*, 2025.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

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

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

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

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

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

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Chunwei Wang, Guansong Lu, Junwei Yang, Runhui Huang, Jianhua Han, Lu Hou, Wei Zhang, and Hang Xu. Illume: Illuminating your llms to see, draw, and self-enhance. *URL https://arxiv.org/abs/2412.06673*, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

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

Shuai Wang, Liang Li, Yang Chen, Ruopeng Gao, Yao Teng, and Limin Wang. Uniddt: Unifying multimodal understanding and generation with decoupled diffusion transformer. *arXiv preprint arXiv:2606.16255*, 2026.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024c.

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

Haiquan Wen, Tianxiao Li, Zhenglin Huang, Yiwei He, and Guangliang Cheng. Busterx++: Towards unified cross-modal ai-generated content detection and explanation with mllm. *arXiv preprint arXiv:2507.14632*, 2025.

Siwei Wen, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Conghui He, Weijia Li, et al. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *Advances in Neural Information Processing Systems*, 38:58972–59005, 2026.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977. IEEE, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Towards instruction-aligned multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Zhonghua Wu, Qingyi Tao, Wentao Liu, Wei Li, and Chen Change Loy. Harmonizing visual representations for unified multimodal understanding and generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17739–17750. IEEE, 2025c.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations*, volume 2025, pp. 28240–28264, 2025.

Zhipei Xu, Xuanyu Zhang, Youmin Xu, Qing Huang, Shen Chen, Taiping Yao, Shouhong Ding, and Jian Zhang. Genshield: Unified detection and artifact correction for ai-generated images. *arXiv preprint arXiv:2605.16122*, 2026a.

Zhu Xu, Jiaqi Tang, Pokai Chen, Yuxin Peng, and Yang Liu. Explainable deepfake detection with feature-robust augmentation and evidence-grounded explanation optimization. *arXiv preprint arXiv:2608.20913*, 2026b.

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*, volume 2025, pp. 70702–70720, 2025.

Chenyang Yang, Shen Yan, Yibo Yang, Litao Hu, Yuchen Liu, Yuan Zeng, Hanchao Yu, Yinan Zhu, Sumedha Singla, Brian Vanover, et al. Detecting ai-generated content on social media with multi-modal language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026)*, pp. 217–229, 2026a.

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

Sen Ye, Mengde Xu, Shuyang Gu, Di He, Liwei Wang, and Winston Hu. Understanding vs. generation: Navigating optimization dilemma in multimodal models. In *International Conference on Learning Representations*, volume 2026, pp. 25118–25141, 2026.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024.

Yanran Zhang, Bingyao Yu, Yu Zheng, Wenzhao Zheng, Yueqi Duan, Lei Chen, Jie Zhou, and Jiwen Lu. D 3 qe: Learning discrete distribution discrepancy-aware quantization error for autoregressive-generated image detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16292–16301. IEEE, 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. *arXiv preprint arXiv:2604.21904*, 2026.

A Appendix

A.1 Baseline Details

Detection Baselines. For explainable synthetic-image detection, we compare with three groups of methods. The first group consists of general-purpose multimodal models, including DeepSeek-VL2 (Wu et al., 2024), InternVL2 (Chen et al., 2024b), Qwen2-VL (Wang et al., 2024b), Qwen3-VL (Bai et al., 2025), Qwen3.5 (Team, 2026), and BAGEL (Deng et al., 2025).

The second group contains specialized synthetic-image detectors: CNNSpot (Wang et al., 2020), FreDetect (Frank et al., 2020), Gram-Net (Liu et al., 2020), Fusing (Ju et al., 2022), LNP (Bi et al., 2023), UnivFD (Ojha et al., 2023), AntiFakePrompt (Chang et al., 2023), FreqNet (Tan et al., 2024a), NPR (Tan et al., 2024b), Fake-HR1 (Jiang et al., 2026), IFM-AIGCSpotter-3B (Yang et al., 2026a), DF-LLaVA (Shen et al., 2025), FatFormer (Liu et al., 2024), AIDE (Yan et al., 2025), SIDA (Huang et al., 2025b), and D^3QE (Zhang et al., 2025).

The third group comprises explainable detectors that generate or exploit natural-language explanations, including FakeVLM (Wen et al., 2026) and UniGenDet (Zhang et al., 2026). UniGenDet serves as the controlled baseline for evaluating IFP. Both models use the same BAGEL initialization, training data, optimization settings, and training budget, differing only in the inclusion of IFP. This comparison evaluates whether incorporating internal flow-error cues improves detection over the existing unified framework. In Table 1, an asterisk denotes a method trained on FakeClue.

Table 7: **Detailed training configurations.** Hardware, optimization, and two-stage training settings for SynerGenDet.

Component	Parameter	Value
Hardware	Training GPUs	8× NVIDIA H200
Optimization	Optimizer	AdamW
	Peak learning rate	1×10^{-4}
	Warm-up steps	2,000
	Adam coefficients	$(0.9, 0.95)$
	Weight decay	1×10^{-2}
	Gradient clipping	1.0
EMA	Decay rate	0.9999
	Updated parameters	All trainable parameters
Stage 1 (IFP)	Detection samples	100K FakeClue
	Generation samples	80K LAION
	Detection:generation ratio	5:1
	Probe timestep	$t_p = 0.5$
	Detection weight	$\lambda_{\text{det}} = 1$
	Explanation weight	$\lambda_{\text{exp}} = 1$
	Flow-matching weight	$\lambda_{\text{fm}} = 1$
Stage 2 (TAA)	Detection expert	Frozen
	Generator feature layer	$K = 8$
	Alignment weight	$\lambda_{\text{TAA}} = 0.5$

Generation Baselines. We compare against generation-only models, including PixArt-Σ (Chen et al., 2024a), Stable Diffusion v2.1 (Rombach et al., 2022), DALL-E 2 (Ramesh et al., 2022), SDXL (Podell et al., 2024), DALL-E 3 (Betker et al., 2023), SD3-Medium (Esser et al., 2024), and FLUX.1-dev (Batifol et al., 2025). We also compare with unified understanding-and-generation models, including TokenFlow-XL (Qu et al., 2025), Janus (Wu et al., 2025a), Emu3-Gen (Wang et al., 2024c), Show-o (Xie et al., 2025), OmniGen2 (Wu et al., 2025b), Janus-Pro-7B (Chen et al., 2025), BAGEL (Deng et al., 2025), and UniGenDet (Zhang et al., 2026).

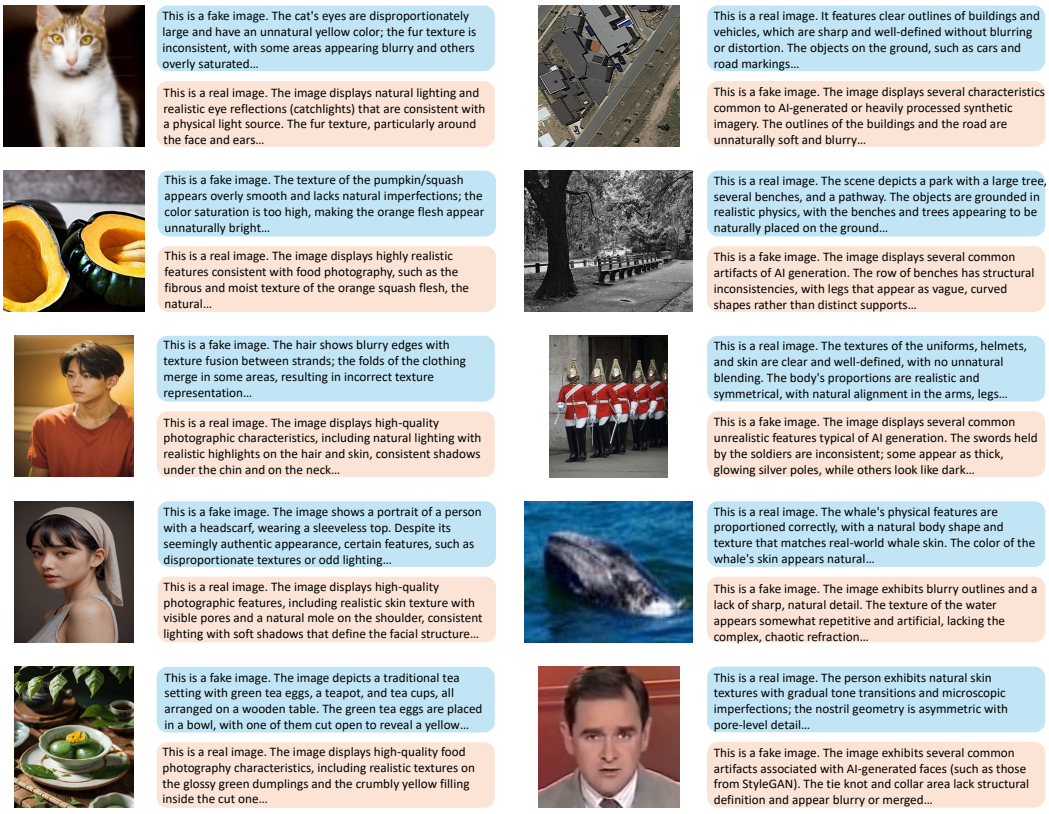


Figure 4: Comparison of detection results. For each sample (left: fake, right: real), the SynerGenDet (top) outperforms the pretrained Qwen3.5-397B-A17B (bottom), providing more accurate detection and superior explanation of artifacts in fake images and features in real ones.

A.2 Implementation Details

Tables 8 and 7 summarize the model architecture and training settings used in our experiments, respectively. Table 8 reports the configurations of the Qwen2.5 language backbone and SigLIP2 vision encoder, while Table 7 provides the optimization hyperparameters and the settings used in the IFP and TAA training stages.

Table 8: **Detailed model configurations.** Specifications of the LLM backbone and vision encoder used in SynerGenDet.

Component	Parameter	Value
LLM Backbone (Qwen2.5)	Hidden size	3,584
	Intermediate size	18,944
	Number of layers	28
	Attention heads	28
	Key/value heads	4
	Vocabulary size	152,064
Vision Encoder (SigLIP2)	Model type	SigLIP2 vision model
	Input image size	980×980
	Patch size	14×14
	Hidden size	1,152
	Number of layers	27

A.3 Qualitative Detection and Explanation Results

Figure 4 presents representative qualitative comparisons between SynerGenDet and the pretrained Qwen3.5-397B-A17B model. Each row contains a synthetic image on the left and a real image on the right. SynerGenDet correctly identifies the authenticity of these samples and supports its predictions with image-specific visual evidence, including unnatural textures, inconsistent object structures, realistic lighting, and natural surface details. In contrast, the baseline predicts the opposite label for these examples and attributes the images to visual cues that are inconsistent with their ground-truth authenticity. These examples illustrate the benefit of combining semantic representations with generative forensic evidence for both authenticity prediction and natural-language explanation.