CLARE: CONTINUOUS LATENT REASONING FOR AIGC IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

As AI-Generated Content (AIGC) advances rapidly, the growing risk of its malicious exploitation creates an urgent need for reliable AIGC image detection. While recent approaches based on Multimodal Large Language Models (MLLMs) have achieved promising detection performance, their reliance on explicit textual chain-of-thought (CoT) introduces two fundamental limitations. First, autoregressively decoding continuous representations into discrete text tokens creates an information bottleneck, compromising rich visual cues. Second, as generated images become increasingly realistic, constructing high-quality and hallucination-free textual reasoning chains becomes substantially more difficult. To overcome these limitations, we propose CLARE, a Continuous Latent Reasoning framework for AIGC image detection. CLARE conducts forensic reasoning entirely within the continuous latent space of MLLMs by directly feeding the last hidden state of each step as the input representation for the next step, thereby bypassing discrete text token generation. Furthermore, we pretrain a task-specific vision encoder on AIGC detection and inject its representations into the latent reasoning process, effectively equipping the general multimodal backbone with forensic sensitivity. Extensive experiments across multiple challenging benchmarks demonstrate that CLARE achieves superior detection performance and exhibits strong cross-generator generalization.

1 INTRODUCTION

Advances from generative adversarial networks and autoregressive text-to-image models to latent diffusion and diffusion transformers have substantially improved the photorealism, semantic alignment, and scalability of image generation (Goodfellow et al., 2014; Ramesh et al., 2021; Ho et al., 2020; Rombach et al., 2022; Saharia et al., 2022; Peebles & Xie, 2023). As a result, synthetic images now exhibit coherent scenes, plausible object structures, and high-quality local details, narrowing the perceptual gap with camera-captured content (Zhu et al., 2023; Chen et al., 2025a; Li et al., 2026b). While this progress expands the scope of visual content creation, it also lowers the barrier to producing convincing synthetic media, raising concerns for content authenticity, digital forensics, and online trust. Reliable AI-generated image detection has therefore become essential, particularly under open-world conditions where new generators and synthesis approaches continue to emerge (Yan et al., 2025; Roca et al., 2026).

Early AIGC image detectors primarily relied on traces left by image generation pipelines, including spatial artifacts, frequency irregularities, gradient patterns, reconstruction residuals, texture inconsistencies, and upsampling traces (Wang et al., 2020; Qian et al., 2020; Tan et al., 2023; Wang et al., 2023; Zhong et al., 2023; Tan et al., 2024). Later studies improved cross-generator transfer by adapting representations from pretrained vision-language models and reducing biases tied to specific training distributions (Ojha et al., 2023; Liu et al., 2024; Tan et al., 2025; Yan et al., 2024; Guillaro et al., 2025). Motivated by the strong reasoning and cross-domain generalization capabilities of multimodal large language models, recent studies have applied them to AIGC detection (Zhou et al., 2025; Wen et al., 2026; Lin et al., 2026; Xu et al., 2026). The resulting models achieve state-of-the-art performance across standard benchmarks and evaluations involving recently released generators, highlighting the promise of MLLMs for generalizable AIGC detection.

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.

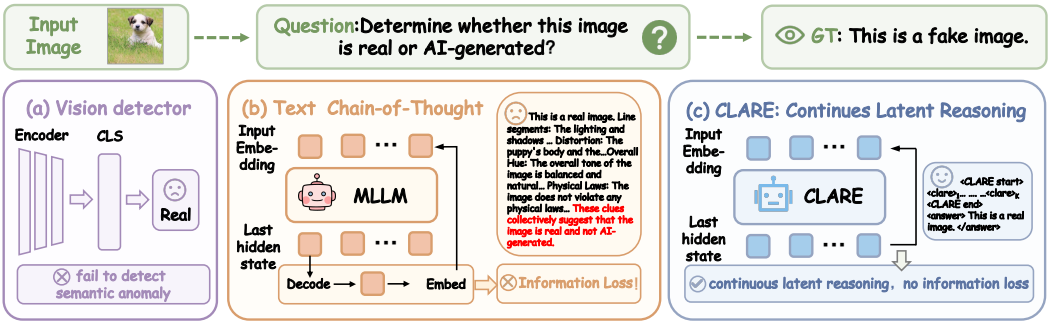


Figure 1: Comparison of three categories of AIGC image detectors. **(a)** Vision-based detectors rely on low-level features and fail to detect semantic-level forgeries. **(b)** MLLM-based detectors that reason via explicit textual chain-of-thought decode continuous hidden states into discrete text tokens at each step, creating an information bottleneck that compromises subtle visual evidence. **(c)** CLARE performs reasoning entirely in continuous latent space, where each step directly uses the last hidden state as the next input, bypassing discrete token generation and thus avoiding the information loss incurred by verbalizing forensic evidence.

Despite this progress, existing detectors exhibit fundamental limitations, as illustrated in Figure 1. Vision-based detectors depend on low-level features and fail to detect semantic anomalies. Many MLLM-based detectors implement their reasoning through explicit textual chain-of-thought (Zhou et al., 2025; Wen et al., 2026; Xu et al., 2026), which may not be an ideal formulation for AIGC detection. First, autoregressive text generation repeatedly maps a high-dimensional hidden state to a discrete vocabulary token and then feeds the corresponding token embedding back into the model. This decode-and-embed process creates an information bottleneck through which subtle visual evidence that is difficult to verbalize may be discarded (Hao et al., 2024; Shen et al., 2025; Jeon et al., 2026; Li et al., 2026a). Such loss is particularly consequential for AIGC detection, which is sensitive to subtle visual traces. Second, textual reasoning requires detailed and visually grounded forensic rationales. As generation quality improves and artifacts become less conspicuous, obtaining such annotations at scale becomes increasingly demanding, while imperfect rationales may introduce unsupported or incomplete evidence into training (Zhou et al., 2025). These limitations may increasingly constrain CoT-based detectors as emerging generators produce more realistic images with subtler and harder-to-verbalize traces.

To address these limitations, we propose **CLARE**, a **C**ontinuous **LA**tent **RE**asoning framework for AIGC image detection. Rather than decoding intermediate states into text, CLARE performs a fixed sequence of reasoning steps directly in the continuous latent space. At each step, CLARE directly uses the last hidden state as the input embedding for the next reasoning step, without projecting it into the vocabulary or re-embedding a predicted token. After completing the latent reasoning, the model returns to text generation and produces the final authenticity judgment. By replacing explicit textual CoT with continuous latent reasoning, CLARE avoids the decode-and-embed bottleneck and eliminates the need for intermediate textual rationales.

However, without explicit supervision, latent states may drift over successive reasoning steps. CLARE therefore introduces step-wise visual soft guidance, which aligns the latent states with spatial visual representations and guides the model to progressively examine the image from left to right and top to bottom. Meanwhile, the vision encoder of a general-purpose MLLM is optimized for general visual understanding rather than synthesis-artifact recognition. We therefore fine-tune a copy of the vision encoder on a real/fake classification objective and inject its region-level forensic representations into the corresponding latent steps, enhancing the model's sensitivity to artifact-related evidence throughout reasoning. After all regions have been processed, the final latent state integrates the evidence accumulated along the reasoning trajectory to predict image authenticity.

Our contributions are summarized as follows:

- We introduce continuous latent reasoning into AIGC image detection, replacing explicit intermediate text with a latent trajectory that avoids the repeated decode-and-embed bottleneck and removes the need for textual CoT supervision.

- We develop a visually grounded latent reasoning framework that combines step-wise spatial alignment with artifact-aware forensic token injection, enabling each reasoning step to integrate regional visual evidence with task-specific forensic features.
- While requiring no textual CoT supervision and generating only a few text tokens at inference, CLARE achieves state-of-the-art performance across multiple challenging AIGC detection benchmarks and maintains consistently high accuracy across diverse forgery granularities and a broad range of generative models.

2 Related Work

2.1 AIGC Image Detection

AIGC image detection aims to distinguish synthesized images from camera-captured ones. Early work mainly exploited low-level traces left by generation pipelines. CNNSpot (Wang et al., 2020) trained a convolutional classifier on images produced by multiple generators (Zhu et al., 2023), and subsequent studies captured synthesis cues from complementary views, such as frequency-domain patterns (Qian et al., 2020), gradient-based representations (Tan et al., 2023), reconstruction residuals (Wang et al., 2023), texture patches (Zhong et al., 2023), and upsampling operations (Tan et al., 2024). Detectors built on such forensic statistics perform well when training and test data share similar distributions, but often degrade when evaluated on unseen generators. UniFD (Ojha et al., 2023) therefore introduced pretrained CLIP representations to obtain more transferable features, and later methods (Chen et al., 2024; Yan et al., 2024; Guillaro et al., 2025) such as FatFormer (Liu et al., 2024) and C2P-CLIP (Tan et al., 2025) further adapted foundation-model representations for cross-generator detection. More recently, several studies have begun to leverage the strong reasoning ability of multimodal large language models for this task. AIGI-Holmes (Zhou et al., 2025) combines semantic visual representations with an artifact-sensitive encoder, GenShield (Xu et al., 2026) jointly learns explainable detection and artifact correction, and Deep-VRM (Lin et al., 2026) injects low-level forensic signals into deeper MLLM layers while preserving early semantic processing. These methods achieve impressive detection performance, demonstrating the potential of multimodal reasoning for AIGC image detection. Nevertheless, methods that produce explicit reasoning traces generally rely on textual chain-of-thought, which faces two limitations. First, converting a continuous hidden state into a discrete text token at every reasoning step incurs information loss. Second, as synthetic images become increasingly photorealistic, collecting fine-grained, evidence-grounded, and hallucination-free textual CoTs becomes more difficult and expensive. In this paper, we propose CLARE to jointly address these limitations through continuous latent reasoning, where the last hidden state of each step directly serves as the input embedding of the next step without generating explicit intermediate text.

2.2 Latent Reasoning

Rather than verbalizing every intermediate step as discrete text, latent reasoning allows a model to perform intermediate computation directly with continuous hidden representations, thereby preserving information that would otherwise be lost during token decoding. This idea was first explored in NLP (Goyal et al., 2024; Deng et al., 2024; Saunshi et al., 2025), where COCONUT (Hao et al., 2024) feeds the last hidden state of each step back as the input of the next to conduct reasoning entirely in continuous space, and CODI (Shen et al., 2025) distills explicit chains of thought into such latent trajectories for more compact reasoning. Building on these advances, recent studies have brought latent reasoning into multimodal large language models by aligning latent states with visual evidence (Jeon et al., 2026; Tong et al., 2025; Wu et al., 2026; Yang et al., 2026). For example, LVR (Li et al., 2026a) trains the model to reconstruct query-relevant visual tokens during reasoning, and Monet (Wang et al., 2026) further optimizes the resulting latent trajectories with a visual-latent policy objective. Similar designs have also proven effective in domain-specific scenarios, as demonstrated by MedLVR (Xi et al., 2026) for medical question answering and Future-L1 (Jiang et al., 2026) for video event prediction. Inspired by these studies, CLARE introduces latent reasoning into AIGC image detection. It aligns latent states with appropriately pooled spatial visual tokens, which avoids both the blurring of artifact-related information caused by excessive pooling and the redundancy introduced by insufficient pooling. Meanwhile, since generic visual features lack sensitivity to synthesis artifacts, CLARE further injects representations from an AIGC-pretrained vision en-

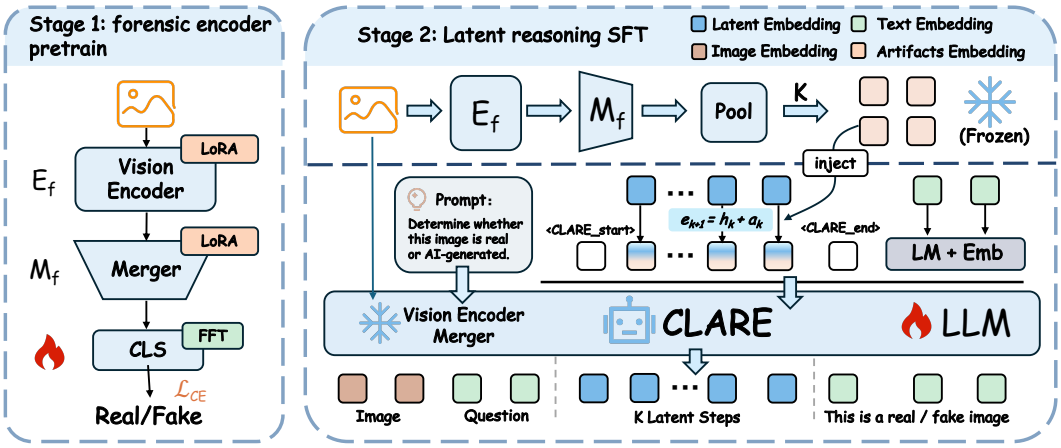


Figure 2: Overview of CLARE. Training proceeds in two stages. Stage 1 pretrains a copied vision encoder and its merger using a real/fake classification objective. In Stage 2, the pretrained forensic encoder is frozen, and its region-level representations are injected into the K continuous latent reasoning steps of the LLM before textual decoding produces the authenticity judgment.

coder into the latent reasoning process, providing the domain-specific forensic knowledge required for reliable detection.

3 METHOD

In this section, we introduce CLARE, a continuous latent reasoning framework for AIGC image detection. As illustrated in Figure 2, CLARE follows a two-stage training pipeline. Stage 1 fine-tunes a copy of the backbone vision encoder for real/fake classification, enhancing its sensitivity to synthesis artifacts. Stage 2 trains the LLM to perform region-wise latent reasoning under spatial visual supervision and incorporates the artifact-aware representations learned in Stage 1. During inference, CLARE performs K latent reasoning steps between <CLARE_start> and <CLARE_end>, after which the LLM resumes standard autoregressive decoding and outputs a textual authenticity judgment for the input image.

3.1 CONTINUOUS LATENT REASONING

An MLLM performs reasoning in an autoregressive manner, where each step builds on the output of the previous one. At step t, the LLM computes a hidden state $\mathbf{h}_t$ from the current input embedding $\mathbf{e}_t$. Textual CoT projects $\mathbf{h}_t$ through the language-model head to predict a token w_t, and the embedding of w_t becomes the input at step $t{+}1$. Latent reasoning removes these token prediction and re-embedding operations, and directly propagates $\mathbf{h}_t$ to the next step (Hao et al., 2024; Jeon et al., 2026). Thus, $\mathbf{e}_{t+1}$ is computed as:

$$\mathbf{e}_{t+1} = \begin{cases} \mathrm{Embed}(w_t), & w_t = \arg\max \mathrm{LM}(\mathbf{h}_t), \quad \text{textual CoT}, \\ \mathbf{h}_t, & \text{latent reasoning}. \end{cases} \tag{1}$$

In CLARE, <CLARE_start> and <CLARE_end> mark the beginning and end of latent reasoning. After generating <CLARE_start>, the LLM performs K consecutive latent reasoning steps, which produce no textual output. Latent reasoning ends at <CLARE_end>, where the LLM returns to token prediction and decodes the final real/fake judgment.

3.2 TRAINING PIPELINE

Stage 1: Forensic encoder pretraining. The vision encoder of a general-purpose MLLM is trained for broad visual understanding and may not respond to the subtle artifacts introduced by generative models. To equip the framework with artifact-level sensitivity, we copy the native vision

encoder as E_{f} and fine-tune it with LoRA (Hu et al., 2021) for binary real/fake classification. A classification head C_{f} is attached to the copied encoder's merger M_{f}, which is optimized together with the LoRA parameters, and the training objective is the cross-entropy loss:

$$\mathcal{L}_{\mathrm{pre}} = -\sum_{c\in\{0,1\}} \mathbb{I}[y=c]\log C_{\mathrm{f}}\big(M_{\mathrm{f}}(E_{\mathrm{f}}(I))\big)_c, \tag{2}$$

where $y \in \{0,1\}$ denotes the ground-truth label with 0 for real and 1 for fake. After pretraining, the classification head is discarded, and the adapted encoder together with its merger is retained as a frozen artifact-aware feature extractor for the next stage.

Stage 2: Latent forensic reasoning. In the second stage, we train the LLM to reason with K continuous states, and reformat each training sample into the structure below:

> ***User:*** *<image> Determine whether this image is real or AI-generated.*
> ***Assistant:*** *<CLARE_start> <clare>$_1$ $\cdots$ <clare>$_K$ <CLARE_end>*
> *<answer> This is a real image. </answer>*

Here, each <clare> is a placeholder occupying one reasoning position, whose input is a continuous hidden state rather than a token embedding.

Step-wise visual alignment. Without explicit supervision, the latent states may drift during the K-step reasoning process. To keep them grounded, we align each latent state with a distinct region of the image. Specifically, the visual tokens $\mathbf{u}$ produced by the native vision encoder's merger are pooled by an operator P_K into K regional representations $\mathbf{V} = P_K(\mathbf{u}) = [\mathbf{v}_1, \ldots, \mathbf{v}_K]$. The hidden state $\mathbf{h}_k$ at step k is then aligned with its corresponding regional representation $\mathbf{v}_k$ through a mean-squared error objective:

$$\mathcal{L}_{\mathrm{latent}} = \frac{1}{K}\sum_{k=1}^{K} \|\mathbf{h}_k - \mathbf{v}_k\|_2^2 . \tag{3}$$

This step-wise alignment assigns each latent step a distinct spatial region to attend to, and guides the K steps to progressively cover the entire image.

Forensic token injection. The visual alignment targets $\mathbf{v}_k$ carry general visual content from the native encoder, but lack sensitivity to synthesis artifacts. To complement this, the forensic encoder's output is pooled with the same operator P_K into $\mathbf{A} = P_K\big(M_{\mathrm{f}}(E_{\mathrm{f}}(I))\big) = [\mathbf{a}_1, \ldots, \mathbf{a}_K]$, where $\mathbf{a}_k$ corresponds to the same spatial region as $\mathbf{v}_k$. At each latent step, the forensic token is fused with the model's hidden state to form the input embedding for the next position:

$$\mathbf{e}_{k+1} = \mathbf{h}_k + \mathbf{a}_k. \tag{4}$$

Latent input construction in training. The latent reasoning steps adopt teacher forcing. Specifically, after the model produces $\mathbf{h}_k$, the input fed to the next step is $\mathbf{v}_k + \mathbf{a}_k$ rather than the predicted hidden state, ensuring the alignment loss directly shapes the target representations learned by the LLM. In this stage, both the native and forensic vision encoders, their mergers, and all other components are frozen. Only the LLM decoder layers are updated.

Training objective. After completing K latent steps, the decoder reaches <CLARE_end>, resumes standard text generation, and produces the answer tokens $\mathbf{w} = (w_1, \ldots, w_T)$. These tokens are supervised by the standard next-token prediction loss:

$$\mathcal{L}_{\mathrm{ntp}} = -\sum_{t=1}^{T} \log p_\theta(w_t \mid \mathcal{C}_{<t}), \tag{5}$$

where $\mathcal{C}_{<t}$ denotes all content preceding w_t. This loss is applied only to the answer tokens, while the K latent steps are supervised solely by the alignment loss $\mathcal{L}_{\mathrm{latent}}$. The overall training objective of this stage is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{ntp}} + \lambda\,\mathcal{L}_{\mathrm{latent}}, \tag{6}$$

where λ balances the two objectives and controls the strength of the visual alignment.

Table 1: Accuracy (%) on Holmes-Set. All existing methods are evaluated with their officially released weights, except Deep-VRM, whose results are taken from the original paper (Lin et al., 2026). **Bold** and underline denote the best and second-best results, respectively.

Method	Janus	Janus-Pro-1B	Janus-Pro-7B	Show-o	LlamaGen	Infinity	VAR	PixArt-XL	SD3.5-Large	FLUX	Mean
CNNSpot	70.00	70.90	85.00	72.20	61.90	86.80	59.90	78.20	63.80	79.90	72.90
AntifakePrompt	72.20	84.30	84.80	86.20	96.20	83.60	90.70	81.70	92.80	66.10	83.90
UniFD	87.60	96.90	96.40	85.90	93.10	79.20	64.30	75.70	87.80	69.60	83.60
NPR	51.20	69.50	73.90	93.70	93.50	93.80	85.90	93.40	91.60	93.60	84.00
LaRE	70.80	74.70	95.60	80.00	91.60	77.90	98.80	82.20	94.10	84.30	85.00
RINE	89.90	98.70	97.20	98.80	99.10	99.20	85.00	98.90	97.80	97.10	96.20
AIDE	91.20	98.90	97.80	98.00	99.40	98.70	93.60	98.60	99.40	94.50	97.00
AIGI-Holmes	95.70	99.10	93.40	97.50	98.00	99.60	99.20	95.20	98.80	79.50	95.60
FakeVLM	86.80	83.30	68.00	78.30	86.80	83.40	82.00	77.10	82.50	83.70	81.20
Deep-VRM	77.40	96.12	96.98	**99.95**	**99.98**	**99.95**	91.45	**99.93**	99.62	**99.67**	96.11
CLARE	**98.24**	**99.32**	**99.70**	98.72	99.82	99.65	**99.72**	99.76	**99.79**	99.64	**99.44**

Table 2: Accuracy (%) on AIGI-Now. All existing methods are evaluated with their officially released weights. **Bold** and underline denote the best and second-best results, respectively.

Method	Nano Banana		GPT-4o		Jimeng		Kling		Minimax		Flux Pro		Flux Krea		Flux Dev		Flux Kontext		AVG
	Pix	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	
NPR	84.4	49.9	91.2	50.1	41.9	49.9	88.3	49.9	52.1	50.0	42.7	50.0	52.3	50.0	89.3	50.0	79.2	49.9	59.5
UniFD	51.0	51.4	53.8	53.3	50.7	52.3	51.4	53.9	50.5	52.1	51.9	51.0	52.2	52.2	63.6	49.9	57.5	55.9	52.3
FatFormer	53.4	49.9	49.9	49.8	48.4	49.9	54.8	50.0	49.1	50.0	48.4	50.0	49.5	50.0	52.4	50.0	63.6	49.9	51.0
SAFE	99.3	50.1	99.7	50.4	50.0	50.5	99.6	50.0	49.8	50.1	49.9	50.5	51.2	50.1	**99.8**	75.3	98.8	83.4	67.1
C2P-CLIP	50.0	50.0	49.8	50.0	49.8	50.0	50.7	49.9	50.3	50.0	52.0	50.0	50.1	50.0	52.6	50.0	60.7	50.0	50.9
AIDE	98.9	51.8	74.7	53.5	63.9	51.4	98.2	55.4	51.4	54.1	60.1	53.8	50.4	56.9	99.1	59.0	97.9	80.6	67.3
DRCT	68.9	57.7	75.0	57.0	89.3	57.7	66.9	56.8	58.7	57.1	83.0	58.2	91.6	58.2	86.5	58.6	86.9	55.5	69.1
AlignedForensics	88.0	50.3	50.8	50.8	82.1	50.0	69.6	50.2	56.7	50.3	65.5	49.9	59.5	50.1	78.6	50.2	65.0	50.0	59.3
CO-SPY	74.1	58.3	78.9	63.1	83.0	76.2	88.3	78.8	77.9	65.6	77.3	72.7	89.1	74.0	88.5	74.9	65.1	66.0	75.1
B-Free	69.6	52.8	59.2	56.7	75.6	54.5	86.1	60.7	56.2	49.9	71.8	54.5	59.5	51.5	74.8	54.5	77.1	53.4	62.2
Deep-VRM	**99.9**	99.0	**100.0**	97.0	99.4	96.5	**99.9**	97.3	93.6	92.4	96.2	97.3	94.4	88.5	99.1	96.3	96.7	79.6	95.7
CLARE	99.7	**99.5**	99.7	**99.4**	**99.6**	**99.4**	99.7	**99.7**	**99.6**	**99.3**	**99.6**	**99.6**	**99.6**	**99.5**	99.6	**99.8**	**99.6**	**98.6**	**99.5**

3.3 Inference

At inference time, the model starts latent reasoning by generating <CLARE_start>. Unlike training, the alignment targets $\mathbf{v}_k$ are no longer used, and the model instead relies on its own hidden state at each step, constructing the next input as $\mathbf{e}_{k+1} = \mathbf{h}_k + \mathbf{a}_k$, where the region-matched forensic representation $\mathbf{a}_k$ continues to provide artifact-aware evidence. After repeating this update for K steps, the model generates <CLARE_end> and produces the final authenticity judgment.

4 Experiments

4.1 Experimental Setup

Evaluation benchmarks. We design three evaluation protocols with distinct focuses. **Protocol 1** adopts the Holmes-Set from AIGI-Holmes (Zhou et al., 2025), which covers ten generators and provides textual chain-of-thought annotations for each sample. **Protocol 2** employs AIGI-Now (Chen et al., 2025a), which covers nine recent commercial and open-source generators, each with pixel-level and semantic-level test subsets that expose different forgery granularities. **Protocol 3** assesses cross-generator generalization on T2I-CoReBench (Li et al., 2026b), which groups generators into Diffusion, Autoregressive, Unified, and Closed-Source categories, MNW (Roca et al., 2026), which collects emerging generators in the wild, and the DRCT benchmark (Chen et al., 2024). We report accuracy throughout. T2I-CoReBench and MNW provide only generated images, so their accuracy is computed over fake samples alone.

Comparison methods. The compared methods span two categories. The first consists of vision-based detectors, including CNNSpot (Wang et al., 2020), UniFD (Ojha et al., 2023), NPR (Tan et al., 2024), FatFormer (Liu et al., 2024), LaRE (Luo et al., 2024), RINE (Koutlis & Papadopoulos, 2024), C2P-CLIP (Tan et al., 2025), AIDE (Yan et al., 2025), DRCT (Chen et al., 2024), Effort (Yan et al., 2024), B-Free (Guillaro et al., 2025), SAFE (Li et al., 2025), AlignedForensics (Sundara Rajan et al., 2025), CO-SPY (Cheng et al., 2025), and DDA (Chen et al., 2025b). The second category

Table 3: Accuracy (%) under Protocol 3. Diff., AR and Closed denote diffusion, autoregressive and closed-source generators. **Bold** and underline denote the best and second-best results.

Method	DRCT	T2I-CoReBench				MNW	Avg.
		Diff.	AR	Unified	Closed		
CNNSpot	49.98	1.16	1.62	1.74	0.61	0.31	17.19
UniFD	61.86	9.03	39.24	27.50	11.45	9.11	30.93
FatFormer	51.99	12.39	61.92	42.62	4.74	12.06	31.49
AIDE	62.56	92.72	99.43	96.41	57.37	29.72	59.59
DRCT	90.59	55.26	75.58	81.20	61.36	57.23	72.06
Effort	69.88	89.20	98.15	91.02	77.97	61.11	73.36
B-Free	**99.21**	72.14	99.08	97.92	62.27	69.29	83.78
DDA	97.99	90.40	99.88	98.64	89.90	82.56	91.75
Deep-VRM	87.51	99.18	99.88	99.28	99.12	86.51	91.13
CLARE	93.16	**99.77**	**99.99**	**99.94**	**99.46**	**96.46**	**96.47**

Table 4: Ablation study on applying CLARE to different backbones. Δ denotes the gain of CLARE over the native Zero-shot backbone. **Bold** indicates the best result.

Backbone	Method	Holmes	AIGI-Now	DRCT	T2I-CoRe	MNW	Avg.
Qwen2.5-3B	Zero-shot	50.00	51.39	6.48	0.15	0.14	21.63
	CLARE	93.28	93.08	91.12	94.79	87.87	92.03
	Δ	+43.28	+41.69	+84.64	+94.64	+87.73	+70.40
Qwen2.5-7B	Zero-shot	57.25	65.26	12.07	31.99	6.84	34.68
	CLARE	97.54	91.68	78.96	96.28	**97.29**	92.35
	Δ	+40.29	+26.42	+66.89	+64.29	+90.45	+57.67
Qwen3-2B	Zero-shot	76.26	77.52	46.46	66.49	36.87	60.72
	CLARE	98.03	98.65	91.01	98.45	96.82	96.59
	Δ	+21.77	+21.13	+44.55	+31.96	+59.95	+35.87
Qwen3-8B	Zero-shot	63.83	70.60	19.70	59.62	22.43	47.24
	CLARE	**99.44**	**99.50**	**93.16**	**99.79**	96.46	**97.67**
	Δ	+35.61	+28.90	+73.46	+40.17	+74.03	+50.43

contains VLM-based and MLLM-based detectors, including AntifakePrompt (Chang et al., 2023), FakeVLM (Wen et al., 2026), AIGI-Holmes (Zhou et al., 2025), and Deep-VRM (Lin et al., 2026).

Implementation details. CLARE is built upon the Qwen3-8B-VL backbone. In Stage 1, the copied vision encoder is fine-tuned with LoRA at rank r=16, scaling factor α=32, and dropout rate 0.05, using a learning rate of 1×10^{-4} with a 0.03 warm-up ratio. In Stage 2, we train the LLM decoder for 2,000 steps on 8 GPUs with a per-device batch size of 2 and gradient accumulation over 8 steps, yielding an effective batch size of 128. The learning rate is 1×10^{-5} with cosine scheduling, the weight decay is 0.1, and the warm-up ratio is 0.03. The latent alignment weight is set to λ=0.1, and the default number of latent reasoning steps is K=36.

4.2 Main Results

Protocol 1: comparison with existing detectors. We first evaluate CLARE on the Holmes-Set against conventional small models and recent MLLM-based detectors across ten generators. As shown in Table 1, CLARE sets a new state of the art with a mean accuracy of 99.44%. It ranks first on five generators and second on four, with Show-o being the only case where it falls outside the top two, and its accuracy stays above 98.2% on all ten generators. Baselines that are competitive on average still fail on individual generators, such as Deep-VRM at 77.40% on Janus and AIGI-Holmes at 79.50% on FLUX. VAR separates the methods most clearly, where baseline accuracies span 59.90% to 99.20% while CLARE reaches 99.72%. Overall, the consistently high accuracy across all ten generators demonstrates the effectiveness of CLARE.

Protocol 2: robustness to forgery granularities. We evaluate CLARE on the pixel-level and semantic-level subsets of AIGI-Now across nine generators. As shown in Table 2, CLARE achieves the best overall result at 99.50% accuracy, ranking first on 14 of the 18 subsets and second on the remaining four. Pixel-level artifacts are generally easier to detect, whereas semantic-level forgeries require identifying inconsistencies in image content rather than relying primarily on low-level traces. CLARE obtains 99.61% accuracy on pixel-level subsets and maintains 99.39% on semantic-level subsets, while many existing detectors degrade substantially on the latter. Its performance remains consistently high across both open-source and commercial generators. These results demonstrate the robustness of CLARE across forgery granularities.

Protocol 3: cross-generator generalization. We evaluate cross-generator generalization on T2I-CoReBench, MNW, and DRCT. As shown in Table 3, CLARE achieves state-of-the-art overall performance under this protocol, with the highest average accuracy of 96.47% across the three benchmarks. It ranks first on all four T2I-CoReBench categories, attaining 99.77%, 99.99%, 99.94%, and 99.46% on the Diffusion, Autoregressive, Unified, and Closed-Source generators, respectively. The 96.46% accuracy on MNW further shows that CLARE remains effective on recently released generators collected in the wild. Overall, CLARE maintains strong performance across diverse generation methods and unseen generator distributions.

Table 5: Comparison among zero-shot, answer-only supervision, textual CoT, and continuous latent reasoning. **Bold** indicates the best result.

Method	Holmes	AIGI-Now	DRCT	T2I-CoRe	MNW	Avg.
Zero-shot	63.83	70.60	19.70	59.62	22.43	47.24
Answer-only	96.90	96.70	81.86	97.36	83.73	91.31
Textual CoT	94.62	96.90	81.11	97.65	94.99	93.05
CLARE	**99.44**	**99.50**	**93.16**	**99.79**	**96.46**	**97.67**

Table 6: Effect of latent reasoning steps without forensic token injection. **Bold** indicates the best result.

K	Holmes	AIGI-Now	DRCT	T2I-CoRe	MNW	Avg.
4	97.48	94.80	81.57	96.09	90.10	92.01
16	97.71	94.61	82.72	97.07	95.11	93.44
36	**99.12**	95.53	**83.93**	**99.09**	**97.42**	**95.02**
64	98.16	**96.04**	83.30	98.30	96.99	94.56

Table 7: Ablation study on the effect of each component. **Bold** indicates the best result.

Variant	Latent reasoning	Visual alignment	Forensic injection	Holmes-Set	AIGI-Now	DRCT	T2I-CoReBench	MNW	Average
Zero-shot				63.83	70.60	19.70	59.62	22.43	47.24
Latent	✓			97.87	95.26	82.96	97.11	93.27	93.29
+ Align	✓	✓		99.12	95.53	83.93	99.09	**97.42**	95.02
CLARE	✓	✓	✓	**99.44**	**99.50**	**93.16**	**99.79**	96.46	**97.67**

4.3 Ablation Study

Generalization across MLLM backbones. As shown in Table 4, we apply CLARE to four backbones from the Qwen3-VL and Qwen2.5-VL families and compare each with its native Zero-shot counterpart. Despite considerable variation in Zero-shot accuracy across backbones, ranging from 21.63% to 60.72%, CLARE consistently raises the average accuracy above 92% for all four, and the improvement is stable across all five evaluation datasets, demonstrating the broad applicability of CLARE across different backbone architectures. Among the four, the Qwen3-8B backbone achieves the best overall performance at 97.67%.

Comparison of reasoning strategies. We compare four variants using the same Qwen3-8B-VL backbone and Holmes-Set training data. **Zero-shot** evaluates the pretrained model without task-specific training. **Answer-only** uses the ground-truth label as the target response, whereas **Textual CoT** is supervised with the CoT annotations provided by Holmes-Set. **CLARE** replaces these textual rationales with continuous latent reasoning. Table 5 reports their average accuracy on five evaluation datasets. The task-specific variants substantially outperform Zero-shot, whose average accuracy is only 47.24%. Answer-only reaches 91.31%, while Textual CoT reaches 93.05%. Textual rationales nevertheless do not consistently outperform direct answer supervision, as Textual CoT improves MNW from 83.73% to 94.99% while falling below Answer-only on Holmes-Set and DRCT. CLARE achieves the best results on all five datasets and raises the overall average from 93.05% for Textual CoT to 97.67%. It improves MNW from 94.99% to 96.46% and DRCT from 81.11% to 93.16%. Overall, CLARE achieves stronger performance than Textual CoT, demonstrating the advantage of continuous latent reasoning for AIGC image detection.

Effect of latent reasoning steps. To evaluate the effect of the number of latent reasoning steps, we disable forensic token injection and vary K across four settings. As shown in Table 6, the average accuracy first rises from 92.01% at $K{=}4$ to 95.02% at $K{=}36$, then falls to 94.56% at $K{=}64$. When K is small, each reasoning step pools visual representations over a large image region, and the coarse aggregation blurs the visual detail needed for accurate detection. Conversely, when K is too large, the image is divided into excessively fine-grained regions, introducing redundancy that limits further improvement. We therefore adopt $K{=}36$, which avoids both the information loss of coarse pooling and the redundancy of over-partitioning.

Effect of each component. To assess the contribution of each component, we successively enable latent reasoning, step-wise visual alignment, and forensic token injection on top of the backbone, as reported in Table 7. Continuous latent reasoning alone achieves an average accuracy of 93.29%. By assigning each latent step to a distinct spatial region, step-wise visual alignment guides the model to examine the entire image systematically and further improves the average accuracy to 95.02%. Adding forensic token injection produces the full CLARE model and achieves the best result on four of the five datasets, increasing the average accuracy to 97.67%. These results verify the effectiveness of each component, and combining them together achieves the best overall performance.

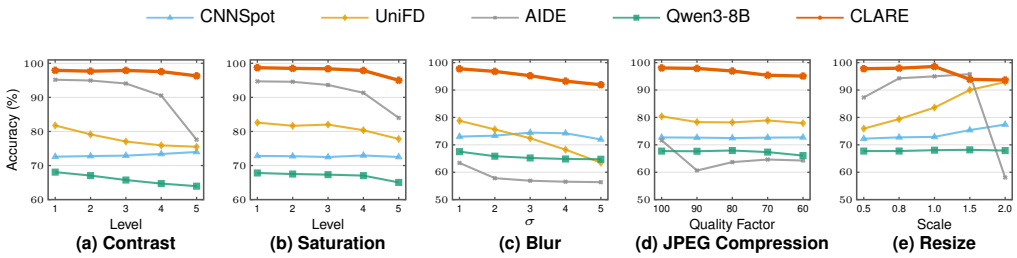


Figure 3: Robustness comparison under five common image post-processing operations.

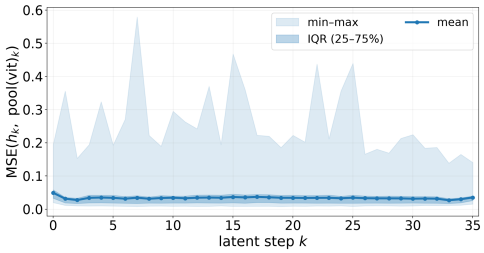


Figure 4: Per-step alignment error $\mathrm{MSE}(\mathbf{h}_k, \mathbf{v}_k)$ between latent hidden states and visual region targets at inference.

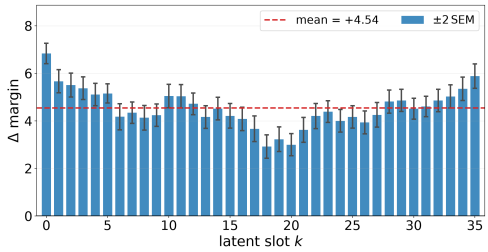


Figure 5: Per-slot discriminative gain (Δmargin) over a zero-latent baseline when only one latent step is injected.

4.4 Robustness Analysis and Visualization

Robustness to image post-processing. We evaluate CLARE against CNNSpot, UniFD, AIDE, and Qwen3-8B with frozen parameters on the same test set under contrast adjustment, saturation adjustment, Gaussian blur, JPEG compression, and resizing. These operations commonly arise during image uploading, sharing, and storage and can weaken subtle synthesis traces. As shown in Figure 3, CLARE maintains consistently strong performance across transformation types and severity levels. In contrast, AIDE is competitive on clean images but degrades sharply under Gaussian blur and JPEG compression, with larger drops as the degradation becomes stronger. The consistently stable performance of CLARE demonstrates its robustness to these post-processing operations.

Visualization of step-wise alignment. Figure 4 shows the alignment error between each latent hidden state and its visual region target at inference. The mean remains near zero with a narrow interquartile range throughout all 36 steps, indicating that the model reliably attends to each assigned region at inference. Figure 5 reports the gain in real/fake decision margin when each slot is injected in isolation, relative to a zero-latent baseline. Every slot yields a positive gain, confirming that each latent step makes an independent contribution to the detection decision, and combining all 36 steps enables CLARE to achieve state-of-the-art performance.

5 Conclusion and Limitations

In this paper, we presented CLARE, a continuous latent reasoning framework for AIGC image detection. Instead of verbalizing forensic evidence as explicit textual chain-of-thought, CLARE propagates the hidden state of each reasoning step directly to the next, keeping the whole reasoning trajectory in the continuous latent space of the MLLM. Each latent step is aligned with a distinct image region and augmented with artifact-aware representations from a task-specific vision encoder, which equips a general-purpose backbone with forensic sensitivity. CLARE's state-of-the-art performance across all three evaluation protocols demonstrates its strong robustness and broad generalizability in AIGC image detection.

A limitation of our approach is interpretability. As reasoning proceeds entirely through continuous hidden states, the intermediate steps are no longer human-readable. Step-wise visual alignment indicates where the model attends, but not what evidence it accumulates. Decoding latent states into explanations on demand is a promising direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with English translation and language polishing of the manuscript, as well as the implementation of the proposed method. All AI-assisted text was repeatedly reviewed and revised by the authors to ensure technical accuracy and consistency with the research. AI-assisted code was manually inspected, executed, debugged, and tested through repeated checks by the authors. The authors made all final decisions regarding the method, experiments, analyses, and claims, and take full responsibility for the content of this paper and the correctness of the implementation.

ETHICS STATEMENT

This work uses only publicly available models and benchmark datasets, some of which may contain images depicting people. No new data were collected or released as part of this study. All datasets are used in accordance with their respective licenses and terms of use. CLARE is intended to support content authenticity and digital forensics. We will release the implementation and model weights under appropriate licenses to facilitate transparent and responsible research use.

REPRODUCIBILITY STATEMENT

Our experiments use the open-source Qwen3-8B-VL model and the Holmes-Set training data from AIGI-Holmes (Zhou et al., 2025), together with publicly available benchmarks for evaluation. The model architecture and training objectives are described in Section 3, while the datasets, implementation details, hyperparameters, and evaluation protocols are provided in the experiment section. Training is implemented with open-source frameworks, including Hugging Face Trainer and DeepSpeed. To support full reproduction of our results, we will release the complete source code, training and evaluation configurations, and model weights.

REFERENCES

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.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *ICML*, volume 2, pp. 6, 2024.

Ruoxin Chen, Jiahui Gao, Kaiqing Lin, Keyue Zhang, Yandan Zhao, Isabel Guan, Taiping Yao, and Shouhong Ding. Task-model alignment: A simple path to generalizable ai-generated image detection. *arXiv e-prints*, pp. arXiv–2512, 2025a.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, et al. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025b.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13455–13465. IEEE, 2025.

Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838*, 2024.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations*, volume 2024, pp. 27896–27923, 2024.

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

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Byungwoo Jeon, Yoonwoo Jeong, Hyunseok Lee, Minsu Cho, and Jinwoo Shin. Vision-aligned latent reasoning for multi-modal large language model. *arXiv preprint arXiv:2602.04476*, 2026.

Tianxiang Jiang, Linquan Wu, Sheng Xia, Songze Li, Ziang Yan, Haoyu Yang, Yu Qiao, and Yi Wang. Imagine before you predict: Interleaved latent visual reasoning for video event prediction. *arXiv preprint arXiv:2606.05769*, 2026.

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

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 148076–148090, 2026a.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 1*, pp. 2405–2414, 2025.

Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? In *International Conference on Learning Representations*, volume 2026, pp. 86729–86758, 2026b.

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

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.

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.

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.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

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

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Thomas Roca, Marco Postiglione, Chongyang Gao, Isabel Gortner, Zuzanna Wojciak, Pengce Wang, Mahsa Alimardani, Shirin Anlen, Kevin White, Juan Lavista Ferres, et al. The microsoft-northwestern-witness benchmark for deepfake detection. *IEEE Intelligent Systems*, 41(2):15–23, 2026.

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.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J Reddi. Reasoning with latent thoughts: On the power of looped transformers. In *International Conference on Learning Representations*, volume 2025, pp. 14855–14881, 2025.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693, 2025.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *International Conference on Learning Representations*, volume 2025, pp. 63304–63322, 2025.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12105–12114. IEEE, 2023.

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, 2024.

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.

Jintao Tong, Jiaqi Gu, Yujing Lou, Lubin Fan, Yixiong Zou, Yue Wu, Jieping Ye, and Ruixuan Li. Sketch-in-latents: Eliciting unified reasoning in mllms. *arXiv preprint arXiv:2512.16584*, 2025.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond image and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12030–12040, 2026.

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.

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.

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.

Linquan Wu, Tianxiang Jiang, Yifei Dong, Haoyu Yang, Fengji Zhang, Shichaang Meng, Ai Xuan, Linqi Song, and Jacky Keung. Lavit: Aligning latent visual thoughts for multi-modal reasoning. *arXiv preprint arXiv:2601.10129*, 2026.

Suyang Xi, Songtao Hu, Yuxiang Lai, Wangyun Dan, Yaqi Liu, Shansong Wang, and Xiaofeng Yang. Medlvr: Latent visual reasoning for reliable medical visual question answering. *arXiv preprint arXiv:2604.09757*, 2026.

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*, 2026.

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.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33510–33520, 2026.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023.

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

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in neural information processing systems*, 36:77771–77782, 2023.

A Overview of Compared Methods

This section summarizes the detectors compared in our experiments.

A.1 Vision-based detectors

CNNSpot (Wang et al., 2020) trains a ResNet-50 classifier on ProGAN images with blur and JPEG augmentation, and shows that a simple convolutional classifier generalizes across GAN architectures when the training data is sufficiently diverse.

UniFD (Ojha et al., 2023) keeps a pretrained CLIP ViT-L/14 frozen and fits only a linear probe on its features, arguing that representations learned outside the forensic domain transfer better to unseen generators than features optimized for a specific generator.

NPR (Tan et al., 2024) targets the upsampling stage shared by most generative pipelines and extracts neighboring pixel relationships as a generator-agnostic artifact descriptor.

FatFormer (Liu et al., 2024) inserts forgery-aware adapters into a frozen CLIP backbone and aligns image patches with text prompts, so that the adapted representation remains anchored to language semantics rather than overfitting to a single training source.

LaRE (Luo et al., 2024) exploits the reconstruction error of a latent diffusion model, using the observation that generated images are easier to reconstruct in latent space than camera-captured ones.

RINE (Koutlis & Papadopoulos, 2024) aggregates intermediate-layer representations of CLIP instead of relying on the final embedding, which preserves low-level cues that later layers discard.

C2P-CLIP (Tan et al., 2025) injects category-related concepts into the CLIP image encoder through prompt-based supervision, improving the separability of real and generated features without modifying the backbone.

AIDE (Yan et al., 2025) combines two ResNet branches over high- and low-frequency image patches with OpenCLIP ConvNeXt-XXL semantic features, so that local texture statistics and global semantics contribute jointly to the decision.

DRCT (Chen et al., 2024) builds diffusion-reconstructed pairs of real images and trains a ConvNeXt-B classifier with contrastive supervision, which makes the decision boundary depend on reconstruction traces rather than on content.

Effort (Yan et al., 2024) decomposes CLIP features by singular value decomposition and preserves the principal subspace while learning a low-rank residual, retaining pretrained knowledge and reducing overfitting to the training generators.

B-Free (Guillaro et al., 2025) constructs self-conditioned inpainted data so that real and fake samples share identical semantic content, and trains a DINOv2 ViT-B detector that is therefore less sensitive to content bias.

SAFE (Li et al., 2025) revisits the preprocessing pipeline and shows that simple choices such as cropping and resizing determine whether generator-specific artifacts survive, achieving competitive accuracy with a lightweight classifier.

AlignedForensics (Sundara Rajan et al., 2025) aligns the training distribution with the target generator through autoencoder reconstruction, allowing the detector to learn artifacts that persist across diffusion variants.

CO-SPY (Cheng et al., 2025) combines semantic features with enhanced artifact features and calibrates their contributions, explicitly addressing the observation that either cue alone degrades on out-of-distribution generators.

DDA (Chen et al., 2025b) performs dual data alignment on both pixel and semantic distributions and fine-tunes a DINOv2 ViT-L/14 backbone with LoRA, reducing the distribution gap between real and generated training samples.

A.2 VLM- AND MLLM-BASED DETECTORS

AntifakePrompt (Chang et al., 2023) formulates detection as visual question answering and tunes soft prompts of a frozen vision-language model, turning authenticity prediction into an answer-generation task.

FakeVLM (Wen et al., 2026) fine-tunes a multimodal large language model to both classify images and describe the artifacts it observes, providing natural-language explanations alongside the decision.

AIGI-Holmes (Zhou et al., 2025) couples semantic visual representations with an artifact-sensitive encoder and supervises the model with human-annotated explanations, so that the produced rationales refer to verifiable visual evidence.

Deep-VRM (Lin et al., 2026) injects low-level forensic signals into the deeper layers of an MLLM while leaving early layers for semantic processing, which equips a general-purpose backbone with low-level features without disrupting its pretrained visual understanding.

B PER-GENERATOR RESULTS

B.1 REASONING-FORM ABLATION ON HOLMES-SET AND AIGI-NOW

Section 3 compares four reasoning forms built on the same Qwen3-8B-VL backbone, and Table 5 reports only their averages over the five evaluation datasets. Here we provide the per-generator breakdown on Holmes-Set and AIGI-Now. **Zero-shot** evaluates the pretrained backbone without task-specific training, **Answer-only** is supervised with the authenticity label alone, **Textual CoT** is supervised with the chain-of-thought annotations released with Holmes-Set, and **CLARE** replaces these rationales with continuous latent reasoning. All four variants share the same training data, backbone, and optimization schedule, so the differences below are attributable to the reasoning form alone.

Table 8: Accuracy (%) of the four reasoning forms on Holmes-Set. All variants use the same Qwen3-8B-VL backbone and training data, and differ only in the form of reasoning supervision. **Bold** and underline denote the best and second-best results, respectively.

Method	Janus	Janus-Pro-1B	Janus-Pro-7B	Show-o	LlamaGen	Infinity	VAR	PixArt-XL	SD3.5-Large	FLUX	Mean
Zero-shot	50.47	68.07	60.42	67.38	96.96	63.58	50.05	68.67	55.84	56.87	63.83
Answer-only	87.52	97.59	96.52	98.78	**99.88**	**99.94**	89.78	**99.91**	99.64	99.41	96.90
Textual CoT	85.88	98.74	98.28	**99.32**	99.60	99.75	66.31	99.71	99.11	99.52	94.62
CLARE	**98.24**	**99.32**	**99.70**	98.72	99.82	99.65	**99.72**	99.76	**99.79**	**99.64**	**99.44**

Table 9: Accuracy (%) of the four reasoning forms on AIGI-Now. *Pix* and *Sem* denote the pixel-level and semantic-level subsets of each generator. **Bold** and underline denote the best and second-best results, respectively.

Method	Nano Banana		GPT-4o		Jimeng		Kling		Minimax		Flux Pro		Flux Krea		Flux Dev		Flux Kontext		AVG
	Pix	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	*Pix*	*Sem*	
Zero-shot	57.6	94.5	53.6	87.9	56.6	92.3	59.6	88.0	54.3	72.7	57.9	85.2	54.7	77.2	64.7	87.3	54.8	71.5	70.6
Answer-only	**99.7**	97.9	**99.9**	93.7	**100.0**	95.8	**100.0**	98.4	**99.7**	95.5	**99.7**	97.5	**99.9**	89.0	**99.8**	98.0	95.5	80.2	96.7
Textual CoT	99.4	98.4	99.5	95.3	99.8	97.6	99.7	97.6	99.2	93.3	99.5	96.2	99.6	90.9	**99.8**	97.3	97.9	82.6	96.9
CLARE	**99.7**	**99.5**	99.7	**99.4**	99.6	**99.4**	99.7	**99.7**	99.6	**99.3**	99.6	**99.6**	99.6	**99.5**	99.6	**99.8**	**99.6**	**98.6**	**99.5**

Table 8 shows that textual chain-of-thought becomes unreliable on the most photorealistic generators. VAR leaves few conspicuous artifacts to describe, and Textual CoT drops to 66.31% there, 23.47 points below Answer-only and close to the 50.05% chance level of the untrained backbone. Janus follows the same pattern at 85.88%. When synthesis traces are hard to put into words, forcing every step through a discrete rationale hurts rather than helps. CLARE keeps the trajectory in continuous latent space and reaches 99.72% on VAR and 98.24% on Janus, and it stays above 98.2% on

all ten generators. Its mean accuracy of 99.44% exceeds Textual CoT by 4.82% and Answer-only by 2.54%, validating the effectiveness of the continuous latent reasoning.

Table 9 reports both forgery granularities of AIGI-Now, and the variants separate mainly on the semantic-level subsets, where detection relies on inconsistencies in image content rather than low-level traces. CLARE is the best variant on all nine of them. It reaches 98.6% on Flux Kontext against 82.6% for Textual CoT and 80.2% for Answer-only, and 99.5% on Flux Krea against 90.9% and 89.0%. Its accuracy also varies little, spanning 98.6% to 99.8% across the nine subsets, whereas the two textual variants swing by more than fifteen points depending on the generator. Averaged over these subsets, CLARE obtains 99.4% against 94.4% for Textual CoT and 94.0% for Answer-only, which directly demonstrates the advantage of continuous latent reasoning for AIGC image detection.

B.2 Per-generator results on T2I-CoReBench

Table 10 reports the accuracy of each compared method on all 40 generator subsets of T2I-CoReBench. Since the benchmark provides only generated images, accuracy equals fake recall. All baselines are evaluated with their officially released checkpoints under their official preprocessing and decision rules, and the overall value is the unweighted mean over the 40 subsets.

C Inference Cost

In practice, an AIGC image detector often has to screen very large collections of images, so the number of tokens generated per image is also as important. Table 11 reports both for the three reasoning forms, using the same Qwen3-8B-VL backbone in all cases. Token statistics are measured over the test samples of the five evaluation datasets.

Textual CoT is by far the most expensive of the three, producing 590 tokens per image on average and up to 1024 on the longest samples. Zero-shot and CLARE are much cheaper, at 6 and 51 tokens. Since the latent trajectory always spans K steps, CLARE emits exactly 51 tokens on every image, so its inference cost stays constant regardless of the input.

A short output is only useful if the decision can be trusted. Zero-shot is the cheapest option but averages 47.24% accuracy, close to random guessing and of little practical value. CLARE reaches 97.67% with about one twelfth of the tokens required by Textual CoT, and is the most accurate variant on all five datasets. Reasoning in latent space thus combines the low cost of a direct answer with accuracy that textual reasoning does not attain, which is what large-scale detection requires.

D Case Study

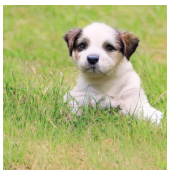
Figure 6 shows a generated image from the Janus subset of Holmes-Set on which the two reasoning forms disagree. The image contains no obvious structural error, and its artifacts are spread over the fur and grass texture rather than concentrated in a namable region. Textual CoT first describes the scene, then examines line segments, edges, texture, shadows, perspective and body structure in turn, and finds each of them natural. Every observation it manages to put into words supports the wrong class, and the verdict follows the rationale to *real*. This is the failure mode behind the low accuracy of Textual CoT on photorealistic generators reported in Table 8. Given the same image, CLARE performs its 36 reasoning steps entirely in latent space, emits no intermediate description, and assigns the fake class a probability of 0.9999.

Table 10: Accuracy (%) on the 40 generator subsets of T2I-CoReBench. All subsets contain only generated images, so accuracy equals fake recall. The Mean row is the unweighted average over all 40 subsets, whereas the T2I-CoReBench column in Table 3 averages the four generator categories. **Bold** indicates the best result on each subset.

Generator subset	CNNSpot	UniFD	FatFormer	AIDE	DRCT	Effort	B-Free	DDA	Deep-VRM	CLARE
BAGEL	1.39	11.57	8.56	89.81	77.55	84.49	92.36	94.44	**100.00**	**100.00**
BAGEL_w_think	0.69	10.19	10.65	96.99	80.09	96.06	95.37	97.92	99.54	**100.00**
BLIP3o-4B	0.00	25.69	36.57	98.38	85.88	76.16	99.77	99.77	96.76	**100.00**
BLIP3o-8B	0.23	30.56	34.49	96.06	78.94	66.20	99.54	99.31	97.69	**100.00**
FLUX.1-Krea-dev	0.23	5.79	5.09	96.30	31.71	87.27	19.91	27.31	98.38	**99.10**
FLUX.1-dev	0.23	4.86	3.70	95.14	52.78	90.74	54.40	94.21	99.54	**99.56**
FLUX.1-schnell	0.69	6.48	11.81	86.11	56.48	92.59	83.10	95.83	98.84	**99.49**
FLUX.2-dev	1.85	4.17	9.26	88.66	36.34	94.91	42.59	88.89	98.61	**99.61**
FLUX.2-klein-4B	1.85	4.40	8.33	84.72	26.62	87.73	31.25	94.21	99.31	**99.91**
FLUX.2-klein-9B	1.62	3.24	3.47	79.86	18.98	82.18	23.61	96.06	**99.54**	99.44
GPT-4o	3.47	6.48	5.09	34.95	2.55	59.95	13.66	73.84	**99.77**	99.00
GPT-Image-1.5	0.69	9.95	1.85	81.71	3.70	91.67	8.56	63.89	99.54	**99.65**
Gemini-2.0-Flash	1.39	24.07	15.51	95.14	82.41	85.19	95.37	95.14	97.22	**97.34**
GoT-R1-7B	2.78	36.81	81.94	99.54	56.71	99.77	98.15	99.77	99.77	**99.98**
HiDream-I1	0.23	3.24	1.16	77.55	67.36	62.96	91.44	82.64	99.54	**99.88**
HunyuanImage-3.0	0.93	22.69	10.88	96.53	86.57	**100.00**	**100.00**	99.77	**100.00**	99.98
Infinity-8B	0.46	41.67	41.90	99.31	94.44	96.53	**100.00**	**100.00**	**100.00**	**100.00**
Janus-Pro-1B	8.10	62.73	88.43	98.38	77.31	**100.00**	97.45	99.77	**100.00**	99.91
Janus-Pro-7B	3.01	40.74	82.64	98.61	74.54	99.77	95.83	99.07	**100.00**	99.98
LongCat-Image	0.00	3.70	9.03	98.61	28.47	94.44	68.52	95.37	**100.00**	**100.00**
Nano-Banana	0.69	6.71	9.03	97.69	79.86	97.92	53.01	81.02	**100.00**	99.24
Nano-Banana-2	0.00	6.48	2.31	93.52	93.52	80.09	38.43	98.38	**99.77**	99.72
Nano-Banana-Pro	0.00	7.18	5.56	95.14	96.30	80.09	71.53	98.61	**100.00**	99.91
OmniGen2-7B	1.39	11.34	23.15	90.74	72.92	89.12	98.84	96.76	98.84	**99.58**
PixArt-Alpha	1.16	23.61	34.72	95.14	78.94	94.21	**100.00**	99.77	99.07	**100.00**
PixArt-Sigma	0.46	24.07	35.42	94.44	77.08	91.90	**100.00**	98.38	99.31	**100.00**
Qwen-Image	0.23	7.41	4.40	97.92	57.87	91.44	96.30	99.07	99.54	**100.00**
Qwen-Image-2512	0.00	10.88	9.95	99.77	46.76	84.49	83.80	99.77	**100.00**	**100.00**
SD-3-Medium	1.62	7.18	13.66	93.98	69.44	93.29	98.61	96.99	99.07	**100.00**
SD-3.5-Large	1.39	18.29	17.82	97.69	71.76	92.36	98.61	98.61	99.54	**100.00**
SD-3.5-Medium	6.25	17.13	28.24	96.06	60.42	97.45	99.07	98.38	99.07	**100.00**
Seedream-3	0.23	4.17	0.00	48.61	45.83	82.18	88.19	94.68	97.92	**100.00**
Seedream-4	0.00	23.84	4.63	10.88	79.40	49.77	95.14	95.14	98.38	**100.00**
Seedream-4.5	0.00	25.93	3.94	41.67	46.06	81.48	98.61	96.76	99.77	**100.00**
Z-Image	0.00	6.71	13.43	96.99	81.02	88.43	97.92	98.38	97.92	**99.65**
Z-Image-Turbo	1.85	2.31	1.16	97.22	77.31	90.05	37.27	72.92	98.84	**99.47**
imagen-4	0.23	5.56	1.39	17.82	69.91	73.38	64.35	95.83	97.92	**99.49**
imagen-4-ultra	0.00	5.56	2.78	13.89	75.46	75.93	58.10	95.60	**100.00**	99.72
show-o2-1.5B	1.16	31.94	45.37	98.61	97.22	98.38	**100.00**	99.54	**100.00**	**100.00**
show-o2-7B	0.46	27.55	85.42	**100.00**	81.02	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Mean	1.17	15.82	20.32	84.25	64.44	87.01	77.22	92.80	99.22	**99.74**

Table 11: Number of tokens generated per image at inference and the corresponding detection accuracy (%). Token counts cover everything the model emits after the prompt, including the latent steps of CLARE. **Bold** indicates the best accuracy.

Method	Generated tokens			Accuracy (%)					Avg.
	Min	Mean	Max	Holmes	AIGI-Now	DRCT	T2I-CoRe	MNW	
Zero-shot	6	6	6	63.83	70.60	19.70	59.62	22.43	47.24
Textual CoT	284	590	1024	94.62	96.90	81.11	97.65	94.99	93.05
CLARE	51	51	51	**99.44**	**99.50**	**93.16**	**99.79**	**96.46**	**97.67**



User: <image> Determine whether this image is real or AI-generated. [GT: Fake]

Textual CoT (predicted *real*, incorrect)

This is a real image.

Image Description. The image depicts a small puppy lying on a grassy field. The puppy has a white coat with brown patches on its ears and around its eyes. It is looking directly at the camera with a curious expression. The grass is green and appears to be well-lit by natural sunlight, casting soft shadows around the puppy.

Based on the provided real image, here are the reasons why this image is real:
1. Line segments. The lighting and shadows in the image are consistent. The puppy and the grass are illuminated by the same light source, and the shadows cast by the puppy align naturally with the direction of the light. There are no unnatural distortions or inconsistencies in the line segments of the grass or the puppy's fur.
2. Edges. The edges of the puppy and the blades of grass are clear and well-defined. There are no discontinuous or inconsistent edges that would suggest AI generation. The pixel distribution around the edges is natural and smooth.
3. Texture. The texture of the puppy's fur and the grass is natural and detailed. There are no repetitive patterns or blurry areas that would indicate artificial generation. The fur appears soft and realistic, and the grass has a natural, varied texture.
4. Distortion. The puppy's body and the surrounding grass do not exhibit any unnatural distortions. The shapes and proportions of the puppy's features are consistent with those of a real puppy.
5. Overall Hue. The overall tone of the image is balanced and natural. The colors are neither overly vibrant nor too dull, and there are no inconsistencies in the hue that would suggest AI manipulation.
6. Clarity. The image has a high resolution with no visible compression artifacts. The details of the puppy's fur and the grass are clear and sharp, with no unnatural pixel blurriness or jagged edges.
7. Perspective. The perspective of the image is consistent and natural. The depth of field changes reasonably, with the puppy in sharp focus and the background slightly blurred, conforming to actual physical laws of optics.
8. Shadows. The shadows in the image are natural and consistent with the light source. The puppy casts a soft shadow on the grass, which aligns with the direction of the sunlight.
9. Physical Laws. The image does not violate any physical laws. The puppy's position on the grass and the interaction between the light and the environment are all consistent with real-world physics.
10. Faces. The puppy's facial features are natural and realistic. There are no signs of overly perfect or distorted features that would indicate AI generation.
11. Body Structure. The puppy's body structure is anatomically correct. The proportions of the head, body, and limbs are consistent with those of a real puppy.
12. Common Sense. The image adheres to common sense. The puppy's pose and the setting are both natural and realistic, with no elements that would be out of place in a real-world scenario.

These clues collectively suggest that the image is real and not AI-generated.

CLARE (predicted *fake*, correct)

<CLARE_start> $\texttt{<clare>}_1$ $\texttt{<clare>}_2$ $\cdots$ $\texttt{<clare>}_{36}$ <CLARE_end>

This is a fake image. $P(\text{fake}) = 0.9999$

Figure 6: A fake image from the Janus subset of Holmes-Set that Textual CoT misclassifies as real. The textual rationale judges every verbalizable aspect of the image to be natural and reaches the wrong verdict, whereas CLARE reasons over 36 latent steps without intermediate text and identifies the image as generated.