CINEMAE: Leveraging Frozen Masked Autoencoders for Cross-Generator AI Image Detection

Minsuk Jang, Hyeonseo Jeong, Minseok Son, Changick Kim

Korea Advanced Institute of Science and Technology

Abstract

While context-based detectors have achieved strong generalization for AI-generated text by measuring distributional inconsistencies, image-based detectors still struggle with overfitting to generator-specific artifacts. We introduce **CINEMAE**, a novel paradigm for AIGC image detection that adapts the core principles of text detection methods to the visual domain. Our key insight is that Masked AutoEncoder (MAE), trained to reconstruct masked patches conditioned on visible context, naturally encodes semantic consistency expectations. We formalize this reconstruction process probabilistically, computing conditional Negative Log-Likelihood (NLL)—$\mathbf{p}(\text{masked} \mid \text{visible})$—to quantify local semantic anomalies. By aggregating these patch-level statistics with global MAE features through learned fusion, CINEMAE achieves strong cross-generator generalization. Trained exclusively on Stable Diffusion v1.4, our method achieves over 95% accuracy on all eight unseen generators in the GenImage benchmark, substantially outperforming state-of-the-art detectors. This demonstrates that context-conditional reconstruction uncertainty provides a robust, transferable signal for AIGC detection.

Introduction

The unprecedented advancement of AI-Generated Content (AIGC), spanning from images and videos to text, has triggered both enormous opportunities and major societal challenges, raising serious legal and ethical concerns. Sophisticated generative models, such as diffusion model (Rombach et al. 2022), now produce synthetic data that is remarkably realistic and widely accessible, enabling creative applications but also facilitating misinformation, copyright infringement, and ethical dilemmas at scales previously unimaginable. Consequently, the development of reliable detectors for synthetic media, especially AI-generated images, has become a pressing research priority.

Initially, supervised deep learning-based detectors dominated the image forensics landscape, relying on large collections of real and synthetic samples for training (Wang et al. 2020; Cai et al. 2023; Wang, Liu, and Wang 2025; Doloriel and Tan 2024). However, these methods tend to overfit to specific generation families (e.g., GANs or diffusion models), resulting in limited generalization when confronted with new or unseen generative paradigms (Corvi et al. 2023). More recent approaches leverage pre-trained,

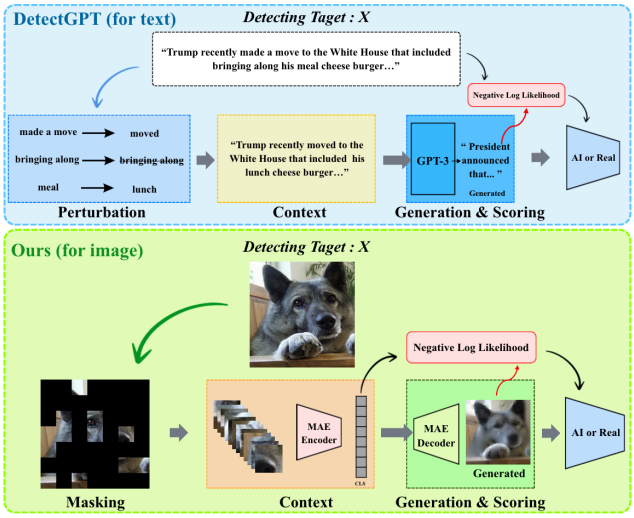


Figure 1: Inspired by context-based text detectors like DetectGPT, our method applies a similar NLL-based paradigm to images. We use a frozen Masked Autoencoder (MAE) to treat unmasked patches as context, reconstructing masked regions and using their conditional NLL for anomaly score.

general-purpose vision backbones (e.g., CLIP (Radford et al. 2021)) or attempt few-shot/zero-shot setups (Ojha, Li, and Lee 2023), yet almost all continue to depend on the right choice of training domain or some prior access to the fake image distribution.

In parallel, the paradigm of context-based detection, pioneered in the text domain with methods like DetectGPT (Mitchell et al. 2023), offers a powerful alternative by measuring content's compatibility with a model's learned distribution. However, translating this to images is challenging. ZED (Cozzolino, Poggi, and Verdoliva 2024a) applies NLL-based detection using a multi-resolution lossless encoder, but it predicts high-resolution pixels conditioned on downsampled, lower-resolution context, thereby losing fine-grained spatial relationships within each resolution level. Conversely, prior works using Masked Autoencoders (MAE) (He et al. 2022) have treated the model as a generic feature extractor for downstream classification (Zhang and Liu 2025). These approaches does not fully

leverage MAE's inherent capability. Building on this observation, we explore whether MAE's reconstruction mechanism can be formalized as a probabilistic generative process for AIGC detection. We propose **CINEMAE** (Contextual INconsistency Estimation with Masked AutoEncoder). To the best of our knowledge, this is the first work to adapt visible patches as context for computing conditional NLL as a principled measure of contextual consistency.

Our main contributions are as follows:

- Inspired by context-based detection in LLMs, we leverage MAE's reconstruction mechanism to adapt visible patches as context for AIGC detection.
- We propose **CINEMAE**, an architecture that synergistically combines local reconstruction uncertainty, derived from MAE's context modeling, with global semantic features from a single backbone for robust, generalized detection.
- We achieve state-of-the-art cross-generator generalization: trained on **Stable Diffusion v1.4 alone, CINEMAE generalizes to all unseen generators** in GenImage with over 95% accuracy, substantially outperforming prior methods

Related Works

Discriminative Feature Learning

Various feature extraction methods have been proposed for AIGC detection. Early approaches like CNNDetection (Wang et al. 2020) relied on low-level artifacts(e.g., upsampling traces, compression fingerprints) to distinguish real from synthetic images. While effective within the same generator family (e.g., GANs), these methods struggled to generalize across different generative paradigms (e.g., GANs to diffusion models). Recent work has shifted toward leveraging pre-trained large-scale models: UnivFD (Ojha, Li, and Eslami 2023) uses CLIP's vision-language feature space for nearest neighbor classification, while reconstruction-based methods like DRCT (Chen, Liu, and Wang 2024) and LaRE (Li, Zhang, and Wang 2024) exploit latent space reconstruction errors. Despite progress, these methods remain fundamentally constrained by their dependence on training data distributions—requiring extensive real-fake pairs and often failing to capture semantic inconsistencies in highly realistic generated content that lack discriminative visual artifacts.

MAE-based Detection Methods

Recent work has adapted Masked AutoEncoders (MAE) (He et al. 2022) for AIGC detection. MARLIN (Cai et al. 2023) applies MAE to facial video representation learning, while MIFAE-Forensics (Wang, Liu, and Wang 2025) and (Doloriel and Tan 2024) explore frequency-domain masking variants for deepfake detection. More recently, (Zhang and Liu 2025) demonstrated that MAE encoder features achieve strong cross-generator performance when paired with a lightweight classifier, shifting focus from low-level artifacts to semantic content.

However, these methods do not leverage MAE's core mechanism—masked image reconstruction from visible context. Instead, they treat MAE as a static feature extractor, essentially borrowing MAE's powerful representation learning capability while discarding the decoder and the masking-reconstruction process entirely. Consequently, they cannot exploit the probabilistic relationship $p(\text{masked} \mid \text{visible})$ learned during pre-training to quantify contextual consistency. Our work addresses this by directly using reconstruction likelihood as a detection signal.

Context-based Detection Methods

DetectGPT (Mitchell et al. 2023) pioneered context-based detection by measuring text perplexity under a language model. ZED (Cozzolino, Poggi, and Verdoliva 2024b) adapted this to images, computing conditional NLL to assess how well images conform to a learned distribution. Specifically, ZED uses a lossless encoder (SReC) that predicts pixels at resolution level ℓ conditioned on context derived from average pooling of the lower-resolution level $\ell - 1$. These pooled patches aggregate surrounding pixel information, and NLL is computed when predicting neighboring pixels from this context. However, this cross-resolution approach introduces limitations: repeated average pooling blurs spatial context, destroying fine-grained patch-level relationships needed for modeling semantic coherence. Additionally, pixel-level modeling limits its ability to capture high-level contextual inconsistencies.

Our method leverages MAE's same-resolution patch context, preserving spatial information within each level. By computing conditional NLL in a semantic feature space rather than pixel space, we maintain the fine-grained relationships critical for detecting contextual anomalies.

Observation and Motivation

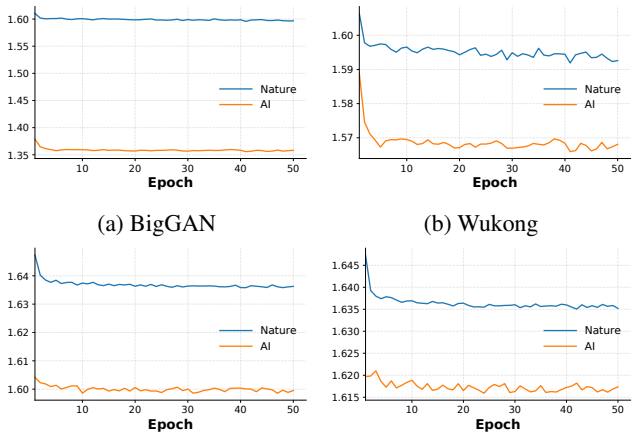
Observation. Our work is motivated by the observation that natural and AI-generated images exhibit distinct statistical properties when evaluated through local context. We hypothesize that the negative log-likelihood (NLL) of image patches, conditioned on their surrounding context, can serve as a discriminative signal for AIGC detection:

$$\text{NLL}(x_m) = -\log p(x_m|x_v; f_\theta) \approx \frac{1}{2\sigma^2}\|x_m - f_\theta(x_v)\|_2^2 \tag{1}$$

where x_m is the masked patch, x_v are visible context patches, and f_θ is the MAE decoder. A high NLL score indicates a local anomaly.

Monitoring NLL scores over 50 training epochs on diverse generators (BigGAN, Midjourney, Stable Diffusion, Wukong—spanning 1 GAN-based, 2 diffusion-based, and 1 mixed dataset), we observe in Figure 2 a striking divergence: AI-generated images converge to near-zero NLL, while natural images plateau at a significantly higher value regardless of continued training. This suggests that generative models learn to produce locally predictable patterns—captured through regularities in frequency domains, upsampling artifacts (Salimans et al. 2016), and smooth local transitions (Ho, Jain, and Abbeel 2020)—which MAE can fully

reconstruct. In contrast, natural images exhibit greater variability in local structure (Zhang et al. 2025), resulting in irreducible reconstruction uncertainty and persistently elevated NLL.



(a) BigGAN (b) Wukong

(c) Midjourney (d) Stable Diffusion v1.4

Figure 2: NLL divergence during training: AI images converge to zero, natural images plateau at higher values.

Limitations of NLL-Only Detection. While NLL clearly separates real from synthetic images, generator-specific variations severely limit cross-domain generalization. As shown in Figure 3, different generators produce fundamentally different NLL distributions due to their distinct architectural biases. To quantify this limitation, we train separate threshold classifiers on each generator in GenImage and evaluate them across all other generators (Figure 4). In-domain performance is strong (e.g., 99.12% on Stable Diffusion v1.4), but cross-domain performance collapses dramatically (e.g., 26.09% on BigGAN when trained on SD v1.4). A single NLL threshold cannot generalize across these domain-specific distributions.

Proposed Solution. This reveals two insights: (1) NLL is highly discriminative within domains, but (2) cross-domain gaps arise from different *types* of predictability across generators, not weak signals. We therefore propose a two-pathway architecture fusing patch-level NLL (generator-specific) with global semantic features (generator-agnostic) to achieve robust cross-domain generalization.

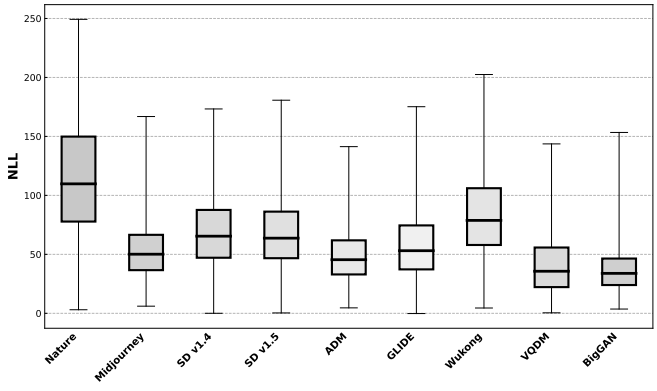


Figure 3: NLL distributions across generators, showing distinct signatures due to architectural differences

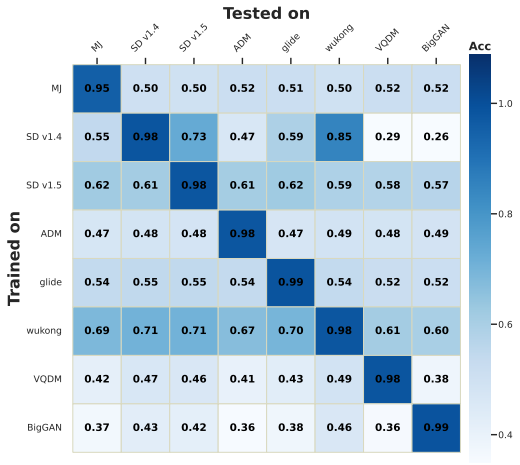


Figure 4: **Cross-domain generalization breakdown.** Each row shows a classifier trained on one generator (e.g., SD v1.4) and evaluated on all generators. Diagonal shows in-domain performance; off-diagonal shows cross-domain failure.

Method

Our preliminary analysis revealed that while local contextual anomaly scores can act as a potent signal for detecting AI-generated images, they often struggle to generalize across different image generators. To overcome this limitation, we introduce **CINEMAE**. CINEMAE is designed to combine both a global semantic representation and a refined local anomaly signal, leveraging the strengths of each to robustly generalize even in zero-shot scenarios. The MAE backbone is kept frozen throughout training, forcing the model to learn generator-agnostic fusion strategies rather than overfitting to any single data source. Figure 5 provides a conceptual overview of our design.

Architectural Components

Semantic Branch (Global Representation). To capture the overall semantic content of the image, this branch extracts the mean of encoder patch features from the frozen MAE encoder. This single vector, denoted as $\mathbf{f}_{\text{global}} \in \mathbb{R}^D$, serves as a robust, high-level representation of the entire image, encapsulating its global composition and semantic structure.

Anomaly Branch (Local Signal). This branch is designed to quantify fine-grained contextual inconsistencies that our preliminary analysis identified as a powerful, albeit non-generalizable, signal. It operates on the MAE's patch reconstruction process under multiple stochastic masks to derive a set of robust anomaly statistics. The core of this branch is our proposed scoring function, described in the follow-

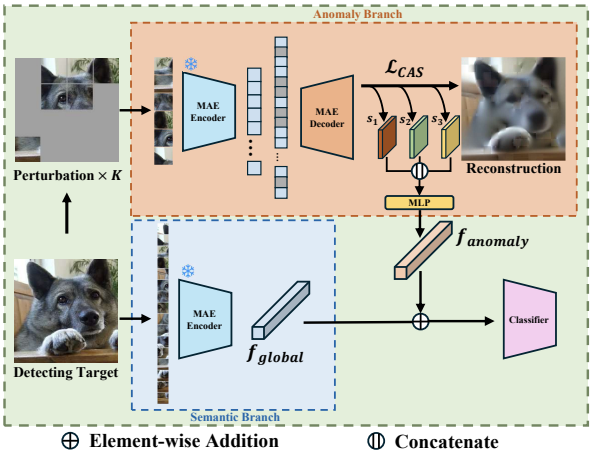


Figure 5: Overview of CINEMAE. Given an input image, we use a frozen MAE encoder to extract a global image representation and a context-based anomaly score. These features are fused and used by a simple classifier to determine whether the image is AI-generated or not.

ing section. Instead of attempting to model a true probability distribution, we propose a novel hybrid scoring function, the *Contextual Anomaly Score* ($\mathcal{L}_{\text{CAS}}$), designed to quantify how "out-of-place" a given patch is. This score is composed of two complementary components: statistical typicality and semantic plausibility.

Local Contextual Anomaly Score. To quantify how "out-of-place" a given patch is, we formulate our Local Contextual Anomaly Score ($\mathcal{L}_{\text{CAS}}$) as a combination of two distinct terms: a statistical deviation term and a semantic mismatch term. This hybrid score allows us to assess a patch's plausibility against both its local context and the model's global understanding.

Formally, for a masked original patch p_i^m, we define its anomaly with respect to the set of visible patches P_v (the context) as:

$$\mathcal{L}_{\text{CAS}}(p_i^m \mid P_v) = \underbrace{\mathcal{D}_{\text{stat}}(p_i^m, P_v)}_{\text{Statistical Deviation}} + \underbrace{\lambda \cdot \|p_i^m - \hat{p}_i^m\|_2^2}_{\text{Semantic Mismatch}} \quad (2)$$

where $\hat{p}_i^m$ is the patch reconstructed by the MAE decoder, and λ is a weighting hyperparameter.

The first term, $\mathcal{D}_{\text{stat}}$, measures the **statistical deviation** of the patch from its visible context. To make this score context-dependent, we explicitly parameterize it using statistics derived from the visible patches P_v. Let $\theta_v = \{\mu_v, \sigma_v\}$ represent the empirical mean and standard deviation of the context P_v. We then formulate the deviation score inspired by the negative log-likelihood (NLL) form, which allows us to quantify the "unlikeliness" of p_i^m given the context's statistics θ_v:

$$\begin{aligned}\mathcal{D}_{\text{stat}}(p_i^m, P_v) &\propto -\log p(p_i^m \mid \theta_v) \\ &= \frac{1}{2}\left(\left(\frac{p_i^m - \mu_v}{\sigma_v}\right)^2 + \log \sigma_v^2\right) \quad (3)\end{aligned}$$

A high value for this term indicates that the patch is statistically improbable given its surroundings.

The second term measures the **semantic mismatch**, penalizing the ℓ_2 distance between the original patch and the one reconstructed by the MAE. A large reconstruction error suggests that the patch violates the image's high-level semantic context.

Crucially, our approach leverages the distinct statistical properties of natural versus AI-generated images. Natural images are known to be statistically complex and non-Gaussian, exhibiting high entropy and long-range correlations that make them inherently difficult to predict from local context (Ruderman 1994; Huang 2000). This results in a naturally higher average NLL when modeled by simple local distributions. Conversely, many AI-generated images, despite their global realism, often contain regions of unnaturally low complexity or statistical regularity, making them more predictable from local context and thus yielding a lower NLL. Our $\mathcal{L}_{\text{CAS}}$ is designed to capture this discrepancy: a high score is characteristic of the complex, unpredictable nature of real-world scenes, while an unnaturally low score can be a powerful indicator of synthetic origin. This allows our model to distinguish between the two domains effectively.

From Anomaly Scores to Image-Level Features

While the patch-level score $\mathcal{L}_{\text{CAS}}$ (Eq. 2) is effective at identifying local inconsistencies, a single high-scoring patch is not sufficient to classify an entire image. An image-level representation must also consider the distribution, magnitude, and stability of these anomalies across the image. Therefore, to form a robust, fixed-size feature vector, we aggregate the $\mathcal{L}_{\text{CAS}}$ scores from all masked patches into a set of three summary statistics. Furthermore, inspired by perturbation-based methods in uncertainty estimation, we repeat this process K times with independent random masks to assess the stability of the anomaly signal.

This yields three complementary statistics s_1, s_2, s_3, where each component is derived from the set of patch-level $\mathcal{L}_{\text{CAS}}$ scores and captures a different aspect of the image's anomalous nature:

- s_1 **(Variability):** The range (max minus min) of the $\mathcal{L}_{\text{CAS}}$ scores across the masked patches, averaged over the K runs. This metric reflects the spatial dispersion of anomalies, distinguishing between concentrated, high-intensity artifacts and more diffuse, subtle inconsistencies.
- s_2 **(Overall Anomaly):** The mean of the $\mathcal{L}_{\text{CAS}}$ scores across all masked patches and all K runs. This provides a global measure of the image's total contextual and semantic implausibility.
- s_3 **(Perturbation Sensitivity):** The mean absolute deviation between each run's average $\mathcal{L}_{\text{CAS}}$ score and the overall mean score s_2. A high s_3 indicates that the anomaly signal is highly sensitive to the choice of context, a characteristic often found in unstructured or chaotic artifacts.

These statistics are then passed through a learned multi-layer perceptron MLP_δ to produce the image-level anomaly feature vector $\mathbf{f}_{\text{anomaly}}$. Collectively, this feature vector $\mathbf{f}_{\text{anomaly}}$

provides a rich description of the image's anomalous properties by capturing not just the average magnitude (s_2) of the $\mathcal{L}_{\text{CAS}}$ scores, but also their distribution (s_1) and stability under perturbation (s_3).

Feature Fusion and Classification

We fuse the global semantic representation $\mathbf{f}_{\text{global}}$ and image-level anomaly feature vector $\mathbf{f}_{\text{anomaly}}$ through additive fusion. This design allows the global representation to be refined by learned anomaly corrections, reflecting the intuition that anomalies provide evidence for revising the baseline global prediction.

After layer normalization, we apply a learned projection MLP_δ to the anomaly features and add the result to the global features:

$$\mathbf{f}_{\text{corrected}} = \mathbf{f}^*_{\text{global}} + \text{MLP}_\delta(\mathbf{f}^*_{\text{anomaly}}) \tag{4}$$

where $(\cdot)^*$ denotes LayerNorm.

Ablation studies (Table 4) demonstrate that this additive strategy outperforms concatenation and learned gating mechanisms, particularly under distribution shift. The advantage stems from its interpretability: rather than requiring the classifier to learn an implicit relationship between disparate feature spaces, additive fusion makes the correction mechanism explicit and parameter-efficient.

Experiments

Experimental Setup

We empirically validate **CINEMAE**'s effectiveness through comprehensive experiments assessing generalization capabilities. **CINEMAE** is trained on a single NVIDIA A100 GPU for 25 epochs with a batch size of 32 using binary cross-entropy loss.

Main Results

We evaluate CINEMAE on the **GenImage** (Zhu et al. 2023b) and **Chameleon** (Wang et al. 2024) benchmarks. We compare against 13 baselines, with results primarily sourced from AIDE (Park, Kim, and Lee 2024). This task is designed to test a detector's ability to identify images from unseen generators, a critical capability.

Generalization Performance on GenImage. As shown in Table 1, CINEMAE achieves state-of-the-art mean accuracy (95.96%), significantly outperforming all 13 models. However, the most critical insight is not just the high average score, but its unprecedented stability across diverse generators. While many state-of-the-art methods like AIDE (86.88%) and PatchCraft (82.30%) show strong performance on some generators, they suffer sharp performance drops on others. For instance, their accuracy drops on GAN-based models like BigGAN. In stark contrast, CINEMAE maintains a consistently high accuracy above 95.9% on every single generator, including those fundamentally different from the training generator (SD v1.4). This remarkable consistency is direct evidence that our method is not merely memorizing artifacts. Instead, by probabilistically modeling the internal contextual consistency of an image, CINEMAE learns a fundamental and generator-agnostic signal of authenticity. This result strongly supports our core hypothesis that a contextual approach, inspired by LLM detecting method, is valid for generalizable AIGC image detection.

Zero-Shot Robustness on Chameleon. We evaluate CINEMAE on the challenging Chameleon benchmark (Wang et al. 2024), a carefully curated dataset of AI-generated images designed to be perceptually difficult to distinguish from real photographs. This benchmark exposes detector biases by revealing that many high-scoring baselines achieve their accuracy through strong bias toward the "real" class, making them unreliable in practice. For example, on the 'All GenImage' split,

Ablation Studies

We conduct a series of ablation studies to validate the core design choices of CINEMAE. All ablation results are model which trained exclusively on SD v1.4 images.

Number of Stochastic Masking (K). We ablate the number of perturbation runs K used to compute s_3. When $K = 0$, only $\mathbf{f}_{\text{global}}$ is used, which yields the lowest accuracy as context-conditional signals are ignored. With $K = 2$, accuracy reaches its peak at an inference time of 41.83ms. While $K = 1$ versus $K = 2$ shows notable performance gains—validating the importance of perturbation sensitivity—comparison with $K = 3$ reveals negligible accuracy improvement despite significantly longer inference time. Since $\mathbf{f}_{\text{anomaly}}$ aggregates mean statistics across K runs, increasing K beyond 2 offers diminishing returns in performance while linearly increasing computational cost. Therefore, $K = 2$ provides an attractive trade-off between detection performance and inference efficiency.

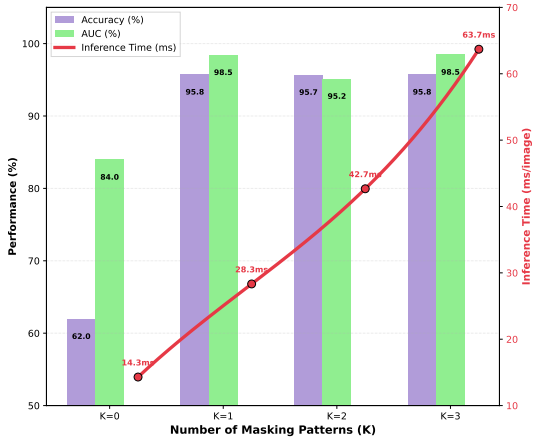


Figure 6: **Ablation on the number of perturbation.** Accuracy peaks at $K = 2$ (41.83ms inference), .

Contribution of Uncertainty Statistics. Each component of $\mathbf{f}_{\text{anomaly}} = [s_1, s_2, s_3]$ is derived from the patch-level $\mathcal{L}_{\text{CAS}}$ scores: s_1 captures spatial variability (range of $\mathcal{L}_{\text{CAS}}$ across patches), s_2 measures overall anomaly magnitude (mean of

Table 1: **Generalization performance on the GenImage benchmark.** We report accuracy (%) on eight unseen generators. All methods are trained only on nature images and AIGCs from a single generator (Stable Diffusion v1.4) to strictly evaluate cross-generator generalization. **CINEMAE** (ours) demonstrates state-of-the-art mean accuracy and, more critically, unprecedented stability across all domains, suggesting it learns fundamental, generator-agnostic signals.

Method	Midjourney	SD v1.4	SD v1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Mean
ResNet-50 (He et al. 2016)	54.90	**99.90**	99.70	53.50	61.90	98.20	56.60	52.00	72.09
DeiT-S (Touvron et al. 2021)	55.60	**99.90**	99.80	49.80	58.10	98.90	56.90	53.50	71.56
Swin-T (Liu et al. 2021)	62.10	**99.90**	99.80	49.80	67.60	99.10	62.30	57.60	74.78
CNNSpot (Wang et al. 2020)	52.80	96.30	95.90	50.10	39.80	78.60	53.40	46.80	64.21
Spec (Zhang et al. 2019)	52.00	99.40	99.20	49.70	49.80	94.80	55.60	49.80	68.79
F3Net (Qian et al. 2020)	50.10	**99.90**	**99.90**	49.90	50.00	**99.90**	49.90	49.90	68.69
GramNet (Liu et al. 2020)	54.20	99.20	99.10	50.30	54.60	98.90	50.80	51.70	69.85
DIRE (Wang et al. 2023)	60.20	**99.90**	99.80	50.90	55.00	99.20	50.10	50.20	70.66
UnivFD (Ojha, Li, and Lee 2023)	73.20	84.20	84.00	55.20	76.90	75.60	56.90	80.30	73.29
GenDet (Zhu et al. 2023a)	89.60	96.10	96.10	58.00	78.40	92.80	66.50	75.00	81.56
PatchCraft (Zhong et al. 2023)	79.00	89.50	89.30	77.30	78.40	89.30	83.70	72.40	82.30
AIDE (Park, Kim, and Lee 2024)	79.38	99.74	99.76	78.54	91.82	98.65	80.26	66.89	86.88
LARE (Luo et al. 2024)	74.00	**100.00**	**99.90**	61.70	88.60	**100.00**	**97.20**	67.80	86.20
CINEMAE (Ours)	**95.90**	98.78	95.92	**96.07**	**95.93**	95.96	**96.01**	**95.90**	**95.96**

Table 2: **Robustness and bias evaluation on the Chameleon dataset.** We report Overall Accuracy (%) and detailed Fake/Real Accuracy (%). To assess in-domain and zero-shot performance, models are trained on the specific splits provided by the dataset. CINEMAE (ours) demonstrates superior fake detection accuracy, avoiding the strong real-class bias exhibited by other state-of-the-art methods.

Training Dataset	CNNSpot	FreDect	Fusing	GramNet	LNP	UnivFD	DIRE	PatchCraft	NPR	AIDE	CINEMAE (OURS)
SD v1.4	60.11 8.86/98.63	56.86 1.37/98.57	57.07 0.00/99.96	60.95 17.65/93.50	55.63 0.57/97.01	55.62 74.97/41.09	59.71 11.86/95.67	56.32 3.07/96.35	58.13 2.43/100.00	**62.60** 20.33/94.38	61.80 **34.47**/81.22
All GenImage	60.89 9.86/99.25	57.22 0.89/99.55	57.09 0.02/99.98	59.81 8.23/98.58	58.52 7.72/96.70	60.42 85.52/41.56	57.83 2.09/99.73	55.70 1.39/96.52	57.81 1.68/100.00	**65.77** 26.80/95.06	62.83 **35.50/84.00**

$\mathcal{L}_{\text{CAS}}$), and s_3 quantifies perturbation sensitivity (mean absolute deviation across random masking). Table 3 demonstrates all three are necessary.

Table 3: **Ablation on anomaly feature vector components.** Full combination (s_1, s_2, s_3) achieves best performance on both GenImage and Chameleon.

Statistics Used			GenImage		Chameleon	
s_1	s_2	s_3	Acc (%)	AUC (%)	Fake	Real
×	×	×	92.21	94.65	28.30	73.21
✓	×	×	94.96	98.11	31.32	76.84
×	✓	×	94.61	99.09	32.77	77.09
×	×	✓	94.66	98.81	34.18	76.72
✓	✓	×	94.81	99.10	34.54	78.15
×	✓	✓	94.73	99.11	33.80	79.53
✓	×	✓	94.78	99.10	34.32	78.71
✓	✓	✓	**95.96**	**99.28**	**34.47**	**81.22**

Feature Fusion Strategy. We validate our choice of fusion strategy. Our architectural hypothesis, detailed in Sec. , posits that treating the local anomaly signal as a context-sensitive correction term to the global representation is the most effective integration method. Table 4 empirically confirms this hypothesis. A additive fusion (Add) significantly outperforms alternatives like concatenation (Concat) or learned gating (Gate). This result suggests that directly allowing the uncertainty signal f_{NLL} to refine the global feature f_{global} is superior to forcing the classifier to learn the complex relationship between these two disparate feature spaces from scratch. The additive approach effectively allows the model to make a baseline prediction using the global context and then "correct" it based on evidence of local inconsistency, a mechanism that proves to be both simple and powerful.

Table 4: **Ablation on feature fusion strategy.** Additive fusion, which treats the uncertainty signal as a correction term to the global feature, yields the best performance.

Method	Early Fusion				Late
	Concat	Gate	Add	Both	Fusion
Accuracy	95.61	94.95	**95.74**	95.52	95.38
AUC	98.94	99.08	**99.28**	99.15	98.75

MAE Freezing Strategy. Freezing both encoder and decoder prevents overfitting to generator-specific artifacts. Figure 7 shows that fully frozen ViT-L/16 achieves the best performance (96.31% average) with minimal variance

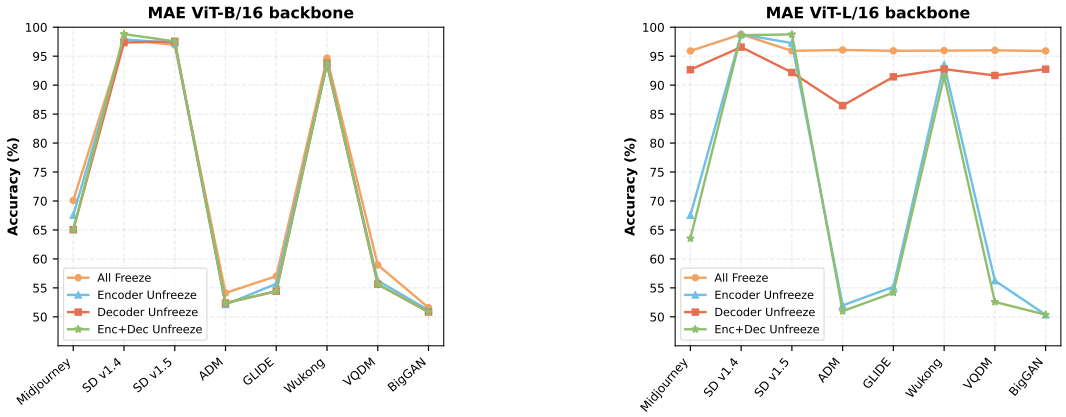


Figure 7: **MAE Freezing Ablation Across Generators.** Left: ViT-L/16 frozen achieves 96.31% average accuracy with minimal variance; unfreezing causes catastrophic collapse. Right: Performance drop (frozen minus unfrozen) for each generator, showing consistent overfitting cost across all models.

(0.94%). Unfreezing any component causes severe overfitting: encoder-only unfreezing drops to 71.36%, decoder-only to 92.06%, and unfreezing both collapses to 70.07% with high variance. ViT-B/16 frozen achieves lower performance (72.66%) due to limited capacity, and unfreezing provides negligible improvement (71.05%), confirming that its constraint is expressiveness, not overfitting. ViT-L/16's frozen superiority demonstrates that large pre-trained features generalize best when protected from fine-tuning.

Conclusion

In this work, we introduced **CINEMAE**, a novel context-based detector for AI-generated images. By probabilistically modeling the relationship between image patches, our method learns a fundamental, generator-agnostic signal of authenticity. Our experiments demonstrate that this approach not only achieves **state-of-the-art generalization** but also avoids the common real-class bias that plagues many existing detectors. Our analysis also reveals a limitation of purely context-based approaches: their performance can degrade in images with low contextual diversity, such as those with uniform backgrounds. We believe this limitation does not diminish our core contribution but instead opens up exciting avenues for future research. Hybrid methods that combine our context-aware signal with frequency-based artifact detectors could offer greater robustness. Furthermore, developing adaptive masking strategies could significantly improve performance in these challenging sparse-context scenarios.

Notably, even without the anomaly features, CINEMAE still demonstrates competitive performance, though not as strong as the full model. This finding suggests that while MAE alone captures strong contextual and representational patterns as a vision foundation model, our additional anomaly-based signals enhance detection capability. This observation implies that **Masked AutoEncoders are latent image detectors.**

References

Cai, Z.; Ghosh, S.; Stefanov, K.; Dhall, A.; Cai, J.; Rezatofighi, H.; Haffari, R.; and Hayat, M. 2023. MARLIN: Masked Autoencoder for Facial Video Representation LearnINg. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1493–1504.

Chen, W.; Liu, X.; and Wang, Z. 2024. DRCT: Discriminative Region-based Counterfeit Tracing for AIGC Detection. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, 7234–7249. PMLR.

Corvi, R.; Cozzolino, D.; Poggi, G.; and Verdoliva, L. 2023. On the Detection of Synthetic Images Generated by Diffusion Models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Cozzolino, D.; Poggi, G.; and Verdoliva, L. 2024a. ZED: A Zero-Shot Entropy Detector for Generative Image Detection. In *Computer Vision – ECCV 2024*, 108–125. Springer Nature Switzerland.

Cozzolino, D.; Poggi, G.; and Verdoliva, L. 2024b. ZED: A Zero-Shot Entropy Detector for Generative Image Detection. In *European Conference on Computer Vision (ECCV)*, 108–124.

Doloriel, C. T. C.; and Tan, K. S. 2024. Frequency Masking for Universal Deepfake Detection. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4780–4784. Seoul, Korea.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 770–778.

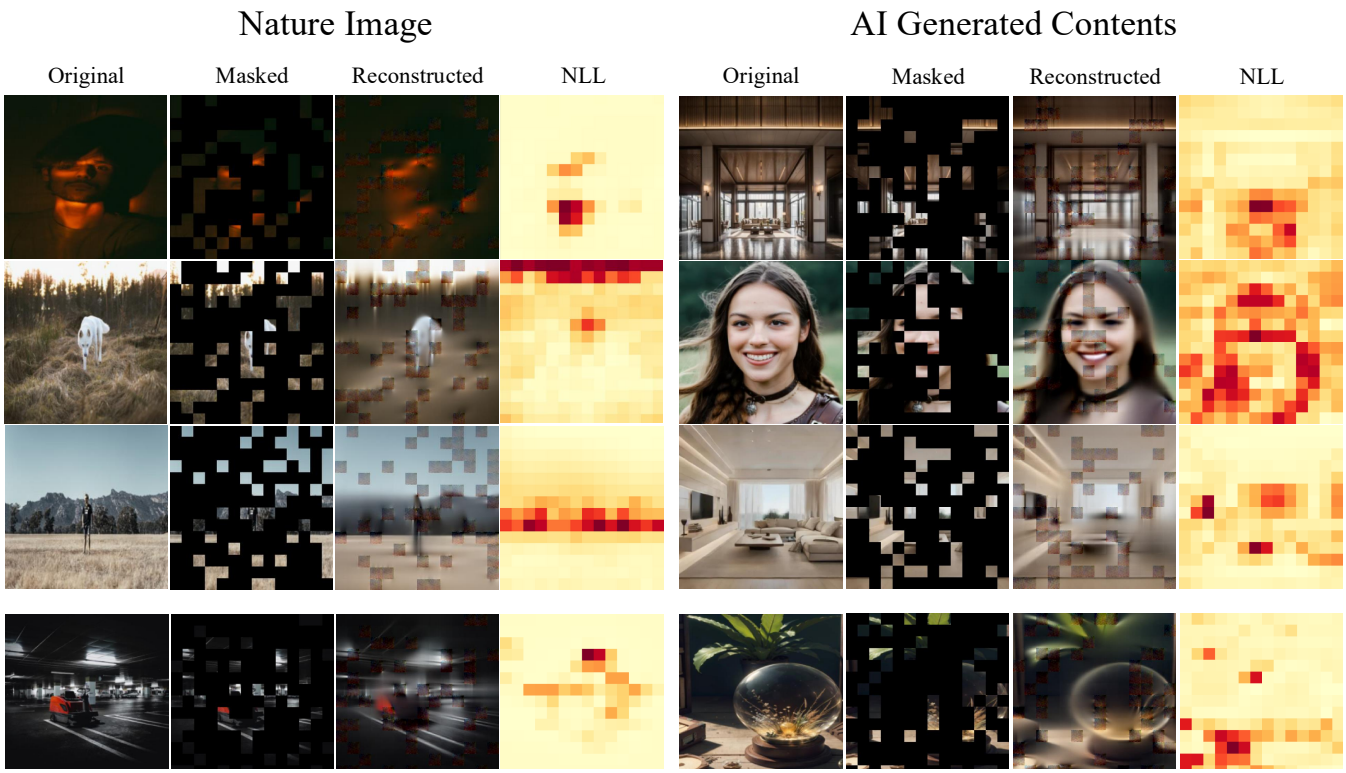


Figure 8: **Qualitative Anomaly Localization on Chameleon.** Per-patch $\mathcal{L}_{\text{CAS}}$ scores (yellow to red indicates increasing contextual inconsistency relative to within-image statistics). CINEMAE successfully identifies implausible regions in both natural and AI-generated images.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*.

Huang, J. 2000. *Statistics of Natural Images and Models*. Ph.D. thesis, Brown University, Providence, RI, USA.

Li, M.; Zhang, Y.; and Wang, Z. 2024. LaRE: Local-aware Representation Enhancement for AI-synthesized Image Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12456–12465.

Liu, Y.-Z.; Li, G.-N.; Wu, Y.-L.; Zhou, Z.-J.; and Zhao, Y. 2020. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 8060–8069.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 10012–10022.

Luo, Y.; Du, J.; Yan, K.; and Ding, S. 2024. LaRE^2: Latent Reconstruction Error Based Method for Diffusion-Generated Image Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7760–7769.

Mitchell, E.; Lee, Y.; Khazatsky, A.; Manning, C. D.; and Finn, C. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, 24584–24598. PMLR.

Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards Universal Fake Image Detectors that Generalize Across Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24480–24489.

Ojha, U.; Li, Y.-J.; and Eslami, S. M. A. 2023. UnivFD: A General-purpose Universal Fake Image Detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24140–24149.

Park, S.; Kim, J.; and Lee, Y. J. 2024. AIDE: Advanced Foundation Model and Discriminative Learning for Fake Image Detection. In *International Conference on Learning Representations (ICLR)*.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision (ECCV)*, 86–103. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume

139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ruderman, D. L. 1994. The Statistics of Natural Images. *Network: Computation in Neural Systems*, 5(4): 517–548.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, 2234–2242.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Jégou, H.; and Tancik, M. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, 10347–10357. PMLR.

Wang, H.; Liu, Z.; and Wang, S. 2025. MIFAE-Forensics: Masked Image-Frequency AutoEncoder for DeepFake Detection. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. Hyderabad, India.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-generated Images Are Surprisingly Easy to Spot... for Now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8695–8704.

Wang, Z.; Bao, J.; Zhou, W.; Chen, W.; Chen, D.; Zeng, W.; and Li, H. 2023. DIRE: A general-purpose dual-mode detector for both natural and AI-generated images. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, Z.; Wu, J.; Xu, W.; Cao, J.; and Wang, Y. 2024. A Sanity Check for AI-generated Image Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, W.; and Liu, Y. 2025. Unmasking Deepfakes: Masked Autoencoding Spatiotemporal Transformers for Enhanced Video Forgery Detection. *IEEE Transactions on Image Processing (TIP)*, 34(3): 567–579.

Zhang, X.; Zhang, X.; Wang, S.-Y.; and Efros, A. A. 2019. Detecting and simulating artifacts in GAN-generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

Zhang, Y.; Nie, J.; Tian, X.; Gong, M.; Zhang, K.; and Han, B. 2025. Detecting Generated Images by Fitting Natural Image Distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhong, H.; Wang, Z.; Ma, Y.; Chen, J.; Zhang, W.; Bao, J.; Zhou, W.; and Li, H. 2023. PatchCraft: A General-Purpose, Patch-Level-Based, and Class-Aware Detector for AI-Generated Images. In *arXiv preprint arXiv:2312.00169*.

Zhu, Y.; Chen, H.; Shen, W.; Liu, Z.; Zhang, Z.; and Liu, W. 2023a. GenDet: A General-Purpose AI-Generated Image Detector. In *arXiv preprint arXiv:2311.08543*.

Zhu, Y.; Chen, H.; Shen, W.; Liu, Z.; Zhang, Z.; and Liu, W. 2023b. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. In *Advances in Neural Information Processing Systems (NeurIPS)*.