ALIGNING BEYOND SEMANTICS: LEARNING DESEMANTICIZED ANCHORS FOR AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

CLIP-based detectors have advanced AI-generated image detection by transferring rich visual representations learned through large-scale image-text pretraining, yet their semantic strengths do not always translate into reliable forgery discrimination. We observe an underappreciated asymmetry on out-of-domain data: the highly competitive performance of a CLIP-adapted detector is driven largely by an exceptionally strong preference for real-image predictions, while its fake-image identification is substantially worse than that of traditional artifact-oriented detectors. To investigate this imbalance, we design a semantic-swap diagnostic that reverses the training-time correlation between image-content semantics and real/fake labels. The evaluated detectors' predictions often follow image content even when it conflicts with an image's real/fake status. We term this Semantic Eclipsing: a failure mode in which semantic cues remain in CLIP's latent space, overshadowing artifact-relevant evidence. In this work, we propose Semantically Unconstrained Alignment (SUA) to suppress semantic eclipsing. Rather than defining alignment targets through fixed semantic descriptions, SUA learns class-specific desemanticized anchors on CLIP's text side through image-anchor contrastive learning. Extensive experiments across multiple benchmarks demonstrate that our method consistently outperforms state-of-the-art approaches, particularly under challenging cross-semantic variations.

1 INTRODUCTION

The rapid evolution of generative models, from generative adversarial networks (GANs) to diffusion-based text-to-image models, has significantly increased the realism and diversity of AI-generated images (AIGI). This progress has intensified the need for generalizable detection methods, particularly under distribution shifts. Many existing approaches exploit generation artifacts through pixel relationships, frequency statistics, or other low-level signals. NPR (Tan et al., 2024b) extracts common up-sampling artifacts present in various generative models. SPAI (Karageorgiou et al., 2025) proposes masked spectral learning to capture spectral artifacts. These artifact-oriented methods mainly capture anomalous patterns within the visual space, providing useful evidence of synthesis. With the rise of diffusion-based text-to-image (T2I) generation, however, leveraging visual-semantic knowledge learned through language supervision offers a complementary direction for detection. CLIP (Radford et al., 2021) provides a rich starting point for such approaches through its representations learned from large-scale image-text pretraining. For example, UnivFD (Ojha et al., 2023) adapts CLIP visual features, significantly improving detection generalization from GANs to unseen diffusion models. DeFake (Sha et al., 2023) concatenates caption-derived CLIP textual features with CLIP visual features to detect T2I images. More recent methods, such as C2P-CLIP (Tan et al., 2025a) and CatAID (Cai et al., 2025a), further exploit semantic knowledge by incorporating caption-based or category-related textual guidance into detector adaptation. Collectively, these methods highlight the potential of CLIP's rich semantic knowledge to improve detection generalization.

However, the aforementioned methods leave the underlying basis of CLIP's advantages insufficiently examined. While artifact-oriented detectors perform classification by identifying artifacts as indicators of synthesis, CLIP-based methods appear to place greater emphasis on recognizing authenticity, exhibiting an *authenticity-oriented* (vs. *artifact-oriented*) tendency. As shown in Fig. 1, we compare

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.

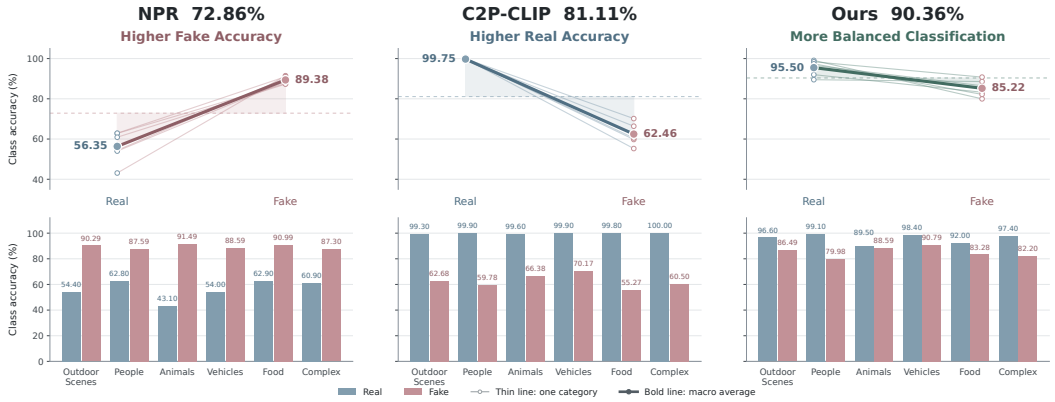

Figure 1: **Contrasting real–fake classification asymmetries** in the artifact-oriented detector NPR and the CLIP-adapted detector C2P-CLIP. Models are trained on UniverForen dataset and evaluated on the out-of-domain CSAIID dataset across six subsets. Top: paired real/fake accuracies; Bottom: per-category accuracies. NPR excels at identifying fake images but is weaker on real images, whereas C2P-CLIP exhibits the opposite tendency. Our SUA improves fake-image identification while retaining high real-image accuracy. Values beside method names report averaged accuracy.

C2P-CLIP with NPR as representative examples of CLIP-adapted and artifact-oriented methods, respectively. Both methods are trained on the full 20-category UniverForen (Ojha et al., 2023) dataset and evaluated on CSAIID (Cai et al., 2025b), where both real and fake images are Out-of-Distribution (OOD). Although NPR excels at identifying fake images, C2P-CLIP achieves higher overall performance owing to its exceptionally high real-image accuracy. This advantage is so pronounced that it overshadows its inferior fake-image accuracy relative to NPR.

Indeed, imbalanced classification behavior reveals a limitation that overall performance alone can obscure: even after task-specific adaptation, a CLIP-adapted detector may not learn sufficiently reliable real-fake distinctions, but instead tends to classify unseen images as real. A natural remedy is threshold calibration to improve classification accuracy. However, current threshold calibration requires prior access to data drawn from the target test distribution (Yang et al., 2026), which may be impractical in real-world detection. This motivates us to investigate the issue at the representation level. Our further analyses (Appendix A.1) reveal an authenticity-oriented feature modeling tendency associated with a *default-to-real* prediction preference in CLIP-adapted detectors.

To investigate the cues governing such preference, we run a diagnostic experiment in a semantic-swapped setting, which reveals the following effect: existing CLIP-adapted methods bias the model toward high-level semantic content, thereby suppressing its sensitivity to low-level forgery artifacts. We formalize this effect as *Semantic Eclipsing* and provide a detailed analysis in Section 3.

To mitigate this issue, in this work, we propose Semantically Unconstrained Alignment (SUA). SUA learns continuous, class-specific textual anchors and uses image-anchor contrastive learning to adapt CLIP visual representations for AIGI detection. These anchors are optimized for detection without predefined linguistic or semantic descriptions, hence termed *desemanticized*. By relaxing static semantic constraints while retaining useful pretrained knowledge, SUA directs discrimination toward forgery-relevant evidence rather than incidental semantic shortcuts. As shown in Fig. 1, this substantially improves fake-image identification while maintaining the strength on real images.

Overall, this work makes the following key contributions:

- We analyse the classification behavior of CLIP-based detectors and reveal their *authenticity-oriented* tendency: exceptionally strong real-image recognition masks inferior fake-image identification.
- We further investigate the origins of this tendency and uncover *Semantic Eclipsing*, an effect in which dominant semantic features in CLIP representations override critical artifact cues.
- We propose Semantically Unconstrained Alignment (SUA), which learns desemanticized textual anchors beyond static semantic constraints while retaining useful pretrained knowledge,

thereby suppressing semantic eclipsing and enhancing identification of forgery artifacts. Our anchor visualization analysis further proves that these generalizable forgery features are tied to desemanticized regions of the language space.

- Extensive experiments across multiple benchmarks demonstrate improved generalization over state-of-the-art methods, particularly under cross-semantic variations. With its flexible plug-and-play design, SUA can be readily integrated into existing VLM-based detectors to further enhance their generalization (Appendix A.2).

2 Related Work

Non-VLM based Detection These methods mostly treat AIGI detection as pure visual binary classification (Wang et al., 2020; Neves et al., 2020; Yu et al., 2022; Tan et al., 2024b). They aim to capture universal artifacts that generalize across architectures and datasets, i.e., the artifact-oriented strategy. A common strategy is to explicitly enhance or isolate forgery artifacts. For instance, CNNSpot (Wang et al., 2020) and NPR (Tan et al., 2024b) exploit upsampling artifacts introduced during image synthesis, while BiHPF (Jeong et al., 2022) amplifies high-frequency components using dual high-pass filtering. LGrad (Tan et al., 2023) leverages gradient-based representations derived from pretrained CNNs to capture subtle inconsistencies. More recently, BiasFree (Guillaro et al., 2025) fine-tunes DINOv2 (Oquab et al., 2024) on diffusion-reconstructed images to isolate generation-specific artifacts. SPAI (Karageorgiou et al., 2025) proposes spectral learning to improve the detection of high-resolution images. While effective, these methods rely solely on visual cues and do not exploit the multimodal representations available in vision-language models.

VLM based Detection Recent work explores VLMs for AIGI detection, leveraging their joint visual-textual representations (Wu et al., 2025; Cai et al., 2025a; Tan et al., 2025a; Yan et al., 2025; Cai et al., 2025b). Existing approaches can be broadly categorized based on how the language modality is utilized. **(1) Text-driven alignment.** These methods introduce text input, such as captions, label words, or prompts, to guide detection through multimodal alignment. For example, LASTED (Wu et al., 2025) formulates detection as an image–text contrastive task using textual labels, while AntifakePrompt (Chang et al., 2023) and FatFormer (Liu et al., 2024) perform prompt learning by inserting learnable tokens into natural language contexts. CatAID (Cai et al., 2025a) and Prompt2Guard (Laiti et al., 2024) further inject semantic categories to enrich textual supervision. These approaches benefit from the semantic knowledge embedded in VLMs, but their reliance on semanticized textual anchors may bias the model toward high-level content. **(2) Vision-only adaptation.** In contrast, another line of work discards the language modality and uses the VLM primarily as a visual feature extractor. UnivFD (Ojha et al., 2023) applies linear probing on CLIP visual features, while Effort (Yan et al., 2025) and ATTSD (Cai et al., 2025b) attempt to disentangle semantic and forgery-related components within the CLIP visual space. Although these approaches seek to mitigate semantic bias, they do not explicitly recognize the authenticity-oriented tendency of CLIP-based detectors and leave the rich latent space of the language modality underexplored. **(3) Chatbot Forensics.** These methods (Tan et al., 2025b; Wen et al.; Huang et al., 2025) usually use Multimodal Large Language Models (MLLMs) and transform the task into multimodal inference for forgery analysis. While these "chatbot-style" methods enrich detection with interpretability, they typically require fine-tuning massive MLLMs with hand-crafted, instruction-following data. Because they mainly focus on forensic analysis rather than generalizability, we view these resource-intensive approaches as a parallel research track, which falls outside the scope of our work.

In summary, existing VLM-based detectors either rely on text input, which may introduce bias, or avoid language altogether, leaving the language embedding space underutilized. This dichotomy motivates us to leverage language representations without inheriting semantic dependencies.

3 Discovery of Semantic Eclipsing

Semantic-Swap Stress Test. To further understand the *authenticity-oriented* behavior of the CLIP-based detectors, we conduct a controlled diagnostic experiment utilizing a semantic-swap stress test. Specifically, we formulate a binary classification task with an intentionally introduced spurious correlation: during training, the real/fake classes are strictly coupled with distinct semantic categories

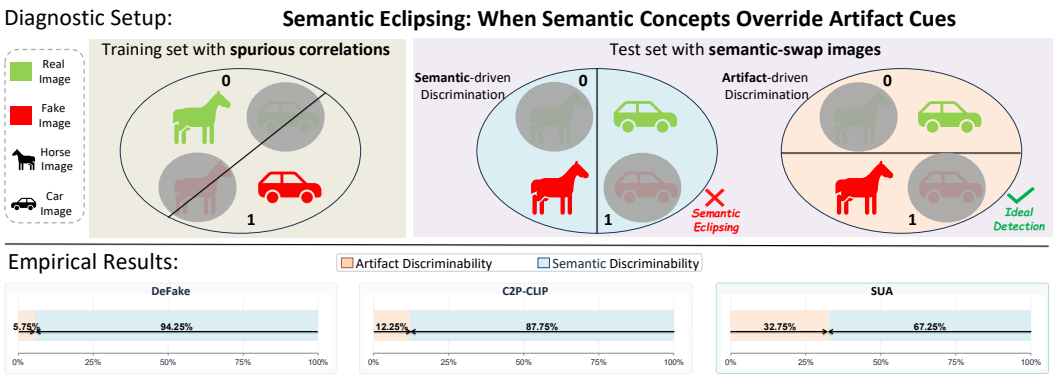

Figure 2: Revealing *Semantic Eclipsing* via intentionally injected spurious correlations. Models are trained on Real-Horse (0) vs. Fake-Car (1) data and perform binary prediction on strictly unseen, *semantic-swap* images. We use two reversed labeling schemes to evaluate two competing modes of image understanding within the model: semantic or artifact discrimination. Semantic and artifact discriminability are measured by test accuracy under object-category labels and real/fake labels, respectively. *Semantic Eclipsing* occurs when the model prioritizes categorical differences to classify images. Compared to baseline methods, our method exhibits the strongest resistance to semantic eclipsing, preserving the highest artifact discriminability.

(e.g., Real-Horse labeled as 0, Fake-Car labeled as 1). At test time, we break this training-time correlation to construct strictly unseen images with swapped pairs (i.e., Fake-Horse and Real-Car). This deliberate semantic-swap setting allows us to evaluate whether the model relies on artifact cues or semantic concepts.

This spurious-correlation setup yields two complementary modes of image understanding: *semantic discrimination* (Car vs. Horse) and *artifact discrimination* (Real vs. Fake: GAN artifacts). To uncover the model's underlying decision mechanism, we measure the artifact discriminability and semantic discriminability by evaluating accuracy under each labeling scheme on strictly unseen test pairs. These two criteria are mutually exclusive: an artifact-sensitive model should show high artifact discriminability, whereas a semantically biased model should show better object recognition but lower artifact discriminability.

As shown in Fig. 2, when subjected to this highly challenging stress test, baseline CLIP-based detectors (DeFake and C2P-CLIP) achieve overwhelmingly high semantic discriminability (up to 94.25%) but fail almost completely in artifact discriminability (dropping as low as 5.75%). This stark imbalance indicates that their predictions are entirely overridden by semantic content. We attribute this Semantic Eclipsing effect to the inherent predisposition of pretrained VLMs to prioritize high-level semantic concepts over low-level forensic artifacts.

Overcoming such deliberately injected spurious correlations remains an exceptionally challenging OOD scenario. However, our empirical results reveal a critical insight. By contrasting image features with our desemanticized textual anchors, the eclipsing effect is significantly mitigated. Even under this extreme stress test, SUA manages to recover substantial artifact discriminability (nearly a 3x to 6x improvement over baselines). To survive semantic eclipsing and recover artifact-discriminative features, a robust detector must redirect CLIP adaptation toward a less semantic-reliant representation. This motivates us to seek desemanticized textual anchors in the latent space.

4 Method

4.1 Desemanticized Textual Anchors

We summarize existing CLIP-adapted detectors in Fig. 3 (left). While these approaches leverage CLIP's rich visual representations during task-specific adaptation, we argue that they do not fully overcome the strong semantic inductive bias inherited from pretraining, which can overshadow the learning of artifact-related cues. To mitigate the *Semantic Eclipsing* effect, as shown in Fig. 3

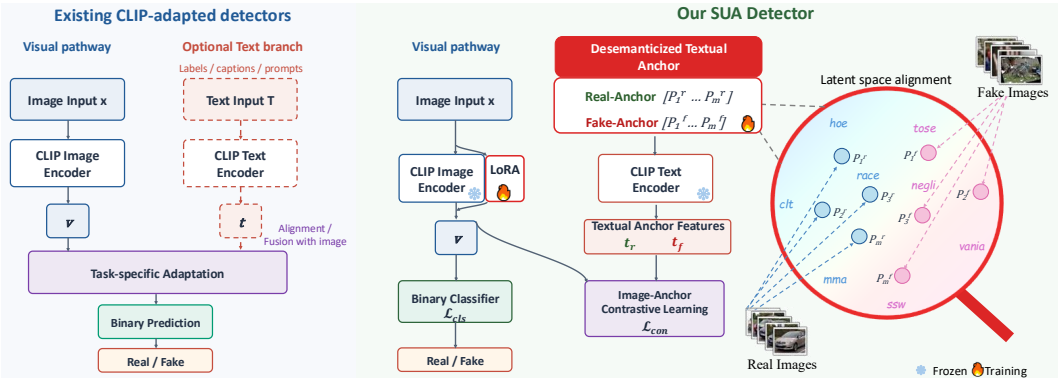

Figure 3: **Existing CLIP-adapted detectors vs. our SUA detector.** Existing methods adapt CLIP visual representations for real/fake classification, optionally incorporating textual guidance, whereas our *Semantically Unconstrained Alignment* (SUA) introduces explicit alignment targets designed to reduce reliance on semantic cues. The magnified view illustrates the alignment between image features and desemanticized textual anchors, depicting how our method maps visual representations to desemanticized regions of the language latent space (the displayed words/phrases and their positions are illustrative; Fig. 4 presents more details).

(right), we learn desemanticized textual anchors within the language space to guide visual adaptation through image–anchor contrastive learning. This design retains CLIP's pretrained language space as a representational foundation for learning effective anchors, while freeing them from static semantic constraints. It thereby reduces reliance on semantic shortcuts during visual adaptation and encourages the discovery of intrinsic forgery patterns.

4.2 Semantically Unconstrained Alignment

Formulation. Given an input image $x \in \mathcal{X} \subset \mathbb{R}^{H\times W\times 3}$, the goal of AIGI detection is to learn a mapping function $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ that assigns a binary label $y \in \{0, 1\}$, where $y = 0$ denotes a real image and $y = 1$ denotes a fake image. Following this formulation, our SUA incorporates CLIP-style image–text contrastive learning. Denoting CLIP's vision and text encoders as $\mathcal{E}_v$ and $\mathcal{E}_t$, respectively, SUA aligns the visual feature $\boldsymbol{v} = \mathcal{E}_v(x)$ with the corresponding class-specific anchor feature $\boldsymbol{t}_y = \mathcal{E}_t(T^y)$, replacing conventional text inputs in CLIP with *desemanticized textual anchors*.

Construction of Desemanticized Textual Anchors. Specifically, we discard the discrete word embedding lookup process. Instead, we define a set of continuous, learnable tokens $P \in \mathbb{R}^{C\times M\times D}$, where $C = 2$ corresponds to the two classes (real and fake), M is the length of the token sequence, and D is the embedding dimension of the text encoder. For each class $c \in \{0, 1\}$, the desemanticized textual anchor T^c is constructed by concatenating the fixed Start-Of-Sequence (SOS) embedding, the learnable tokens P^c, and the fixed End-Of-Text (EOT) embedding:

$$T^c = \left[\boldsymbol{e}_{\text{SOS}}; P^c; \boldsymbol{e}_{\text{EOT}}\right] \in \mathbb{R}^{(M+2)\times D}. \tag{1}$$

Image–Anchor Contrastive Learning. The semantically unconstrained alignment is performed via image–anchor contrastive learning. The input image x is fed into a vision encoder $\mathcal{E}_v$ adapted via LoRA (Hu et al., 2022) tuning to obtain the visual feature $\boldsymbol{v} = \frac{\mathcal{E}_v(x;\theta_{\text{LoRA}})}{\|\mathcal{E}_v(x;\theta_{\text{LoRA}})\|_2}$. The anchor T^c is then fed into the frozen text encoder $\mathcal{E}_t$ to obtain the textual anchor features $\boldsymbol{t}_c = \frac{\mathcal{E}_t(T^c)}{\|\mathcal{E}_t(T^c)\|_2}$. The visual feature of each image with label y forms a positive pair with the corresponding class anchor $\boldsymbol{t}_y$ and a negative pair with $\boldsymbol{t}_{1-y}$.

Loss Function. The primary supervision comes from the contrastive loss, which aligns the visual feature $\boldsymbol{v}$ with the text embeddings of the corresponding class anchor $\boldsymbol{t}_y$:

$$\mathcal{L}_{con} = -\log \frac{\exp(\tau\langle \boldsymbol{v}, \boldsymbol{t}_y\rangle)}{\sum_{c=0}^{1}\exp(\tau\langle \boldsymbol{v}, \boldsymbol{t}_c\rangle)} \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity and τ is a learnable temperature scalar. To further stabilize training, we introduce a classification head $h(\cdot)$ on top of the visual features. This head is optimized using the Binary Cross-Entropy (BCE) loss:

$$\mathcal{L}_{cls} = \text{BCE}(h(\boldsymbol{v}), y) \tag{3}$$

The final optimization objective is thus defined as $\mathcal{L}_{total} = \mathcal{L}_{con} + \lambda \mathcal{L}_{cls}$, where λ is a hyperparameter balancing the two loss terms.

5 Experiments

5.1 Implementation Details

We utilize the pretrained CLIP ViT-L/14 as the backbone model and use Adam optimizer with a learning rate of 4×10^{-4} and a batch size of 128. For the LoRA parameters, following C2P-CLIP (Tan et al., 2025a), we set the bottleneck rank $r = 6$, the scaling factor $\alpha = 6$, and the dropout probability to 0.8. All experiments were conducted using PyTorch on a server with four NVIDIA RTX A6000 GPUs (48GB each). Note that the text encoder is only used for supervising the training phase and will be discarded during inference. The number of tokens for each class is empirically set to 16. The hyperparameter λ is set to 8.0. For more details, please see the Appendix A.3.

5.2 Evaluation Protocols and Datasets.

Cross-Generator Generalization. We first evaluate our method on widely used single-source generalization benchmarks: UniverForen (Ojha et al., 2023) and GenImage (Zhu et al., 2024). These benchmarks focus on cross-generator generalization, where the training set pairs one real source with one generator. **GAN-Source Generalization:** UniverForen evaluates generalization from a GAN-based generator. The training set contains images from ProGAN across 20 semantic categories, along with real images from the LSUN dataset (Yu et al., 2015). The test set includes images generated by 19 other generators. **Diffusion-Source Generalization:** GenImage focuses on diffusion-source generalization. The training set comprises images generated by Stable Diffusion v1.4 (SD v1.4) paired with real images from ImageNet (Deng et al., 2009). The test set includes multiple generators, including 1 GAN-based model and 7 diffusion models.

Cross-Semantic Generalization. To assess cross-semantic generalizability, following the setting of ATTSD, we test on the Cross-Semantic AI-Generated Image Detection (CSAIID) dataset (Cai et al., 2025b). Its two training protocols follow those of UniverForen and GenImage, but the test set is composed of unseen, diverse categories focusing on broader semantic concepts such as outdoor scenes, people, animals, vehicles, food, and complex scenarios.

Cross-Dataset Generalization. We further establish a cross-dataset protocol based on the recently introduced Community Forensics (CF) dataset (Park & Owens, 2025). The training data of CF comprises images generated by thousands of models, including GANs, diffusion models, and autoregressive generators. We use the small version of the CF training set. The test set is sampled from UniverForen (7 GAN subsets) and the GenImage dataset. Compared to single-source benchmarks, this multi-source setting provides a more realistic scenario for evaluating generalization.

Cross-Resolution Generalization. Following the recent benchmark introduced by SPAI (Karageorgiou et al., 2025), we further evaluate our method's generalization across image resolutions. The evaluation protocol and detailed results are provided in Appendix A.4.

5.3 Compared with SOTA Methods

Cross-Generator Evaluation. The cross-generator generalization performance of our method compared to SOTA approaches on both GAN-based and diffusion-based benchmarks is presented in Table 1 and Table 2, respectively. Our proposed method outperforms all competing methods on both benchmarks, demonstrating its strong generalization capability across a diverse range of unseen generators.

Table 1: Accuracy (ACC, %) comparison of **cross-generator** generalization on the UniverForen dataset. All methods were trained on the ProGAN-generated images vs. real images from LSUN and evaluated on subsets from different generators. * indicates the retrained version of the official code.

Method	Generative Adversarial Networks						Deep fakes	Low level vision		Perceptual loss		Guided	LDM			Glide			DA LLE	Total
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/ CFG	100 steps	100 27	50 27	100 10		Avg. Acc.(%)
CNNSpot (Wang et al., 2020)	99.99	85.20	70.20	85.70	78.95	91.70	53.47	66.67	48.69	86.31	86.26	60.07	54.03	54.96	54.14	60.78	63.80	65.66	55.58	69.58
Patchfor (Chai et al., 2020)	75.03	68.97	68.47	79.16	64.23	63.94	75.54	75.14	75.28	72.33	55.30	67.41	76.50	76.10	75.77	74.81	73.28	68.52	67.91	71.24
Co-occurence (Nataraj et al., 2019)	97.70	63.15	53.75	92.50	51.10	54.70	57.10	63.06	55.85	65.65	65.80	60.50	70.70	70.55	71.00	70.25	69.60	69.90	67.55	66.86
Spec (Zhang et al., 2019)	49.90	99.90	50.50	49.90	50.30	99.70	50.10	50.00	48.00	50.60	50.10	50.90	50.40	50.40	50.30	51.70	51.40	50.40	50.00	55.45
AIDE* (Yan et al., 2024)	99.99	98.48	83.95	99.64	73.25	99.91	52.21	64.17	50.91	80.05	86.74	53.10	81.35	62.90	82.00	64.20	65.20	64.20	71.30	75.45
UnivFD (Ojha et al., 2023)	100.0	98.50	94.50	82.00	99.50	97.00	66.60	63.00	57.50	59.50	72.00	70.03	94.19	73.76	94.36	79.07	79.85	78.14	86.78	81.38
NPR* (Tan et al., 2024b)	99.99	86.83	87.70	96.90	84.49	99.52	66.88	64.44	70.09	50.00	50.00	74.65	90.90	93.15	92.20	95.20	95.95	95.90	87.95	83.30
LGrad(Tan et al., 2023)	99.84	85.39	82.88	94.83	72.45	99.62	58.00	62.50	50.00	50.74	50.78	77.50	94.20	95.85	94.80	87.40	90.70	89.55	88.35	80.28
FreqNet (Tan et al., 2024a)	97.90	95.84	90.45	97.55	90.24	93.41	97.40	88.92	59.04	71.92	67.35	86.70	84.55	99.58	65.56	85.69	97.40	88.15	59.06	85.09
FatFormer (Liu et al., 2024)	99.89	99.32	99.50	97.15	99.41	99.75	93.23	81.11	68.04	69.45	69.45	76.00	98.60	94.90	98.65	94.35	94.65	94.20	98.75	90.86
AntifakePrompt (Chang et al., 2023)	82.79	78.09	81.77	72.92	90.11	74.94	89.47	79.94	65.00	50.00	50.00	78.05	82.60	70.35	82.65	87.00	86.85	86.95	81.70	77.43
CatAID (Cai et al., 2025a)	99.86	98.83	97.00	96.63	96.24	57.28	99.35	71.53	82.22	79.05	79.87	78.80	92.35	88.45	92.50	88.80	89.30	89.25	91.35	87.82
C2P-CLIP* (Tan et al., 2025a)	99.95	99.66	99.02	91.38	99.70	99.80	93.84	89.17	88.13	98.50	98.53	63.80	98.55	92.90	98.70	95.05	95.50	96.00	95.85	94.42
ATTSD (Cai et al., 2025b)	99.80	98.64	93.75	97.86	98.03	100.00	74.26	50.83	53.88	98.12	99.15	89.50	99.20	99.15	99.60	99.45	99.45	99.80	99.60	92.11
Effort (Yan et al., 2025)	100.0	99.85	99.60	95.05	99.60	100.0	87.60	92.50	81.50	98.90	98.90	69.15	99.30	96.80	99.45	97.45	97.80	97.15	98.05	95.19
Ours	100.00	95.00	93.97	98.20	96.11	99.70	85.00	92.92	95.39	95.38	84.30	99.65	99.15	99.80	98.65	98.80	98.60	99.65	92.53	**95.94**

Table 2: Accuracy (ACC, %) comparison of **cross-generator** generalization on the GenImage dataset. All methods were trained on SD v1.4-generated images vs. real images from ImageNet and evaluated on subsets from different generators.

Methods	Venues	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	mAcc
ResNet-50 (He et al., 2016)	CVPR 2016	54.9	99.9	99.7	53.5	61.9	98.2	56.6	52.0	72.1
DeiT-S (Touvron et al., 2021)	ICML 2021	55.6	99.9	99.8	49.8	58.1	98.9	56.9	53.5	71.6
Swin-T (Liu et al., 2021)	ICCV 2021	62.1	99.9	99.8	49.8	67.6	99.1	62.3	57.6	74.8
CNNSpot (Wang et al., 2020)	CVPR 2020	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
Spec (Zhang et al., 2019)	WIFS 2019	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8
F3Net (Qian et al., 2020)	ECCV 2020	50.1	99.9	99.9	49.9	50.0	99.9	49.9	49.9	68.7
GramNet (Liu et al., 2020)	CVPR 2020	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9
DeFake (Sha et al., 2023)	CCS 2023	79.9	98.7	98.6	71.6	78.1	98.4	78.3	74.4	84.7
UnivFD (Ojha et al., 2023)	CVPR 2023	91.5	96.4	96.1	58.1	73.4	94.5	67.8	57.7	79.5
NPR (Tan et al., 2024b)	CVPR 2024	81.0	98.2	97.9	76.9	89.8	96.9	84.1	84.2	88.6
FreqNet (Tan et al., 2024a)	AAAI 2024	89.6	98.8	98.6	66.8	86.5	97.3	75.8	81.4	86.8
FatFormer (Liu et al., 2024)	CVPR 2024	92.7	100.0	99.9	75.9	88.0	99.9	98.8	55.8	88.9
DRCT (Chen et al., 2024)	ICML 2024	91.5	95.0	94.4	79.4	89.2	94.7	90.0	81.7	89.5
CatAID (Cai et al., 2025a)	ICCVW 2025	80.4	98.8	98.8	55.3	90.4	94.0	68.1	70.7	82.1
Effort (Yan et al., 2025)	ICML2025	82.4	99.8	99.8	78.7	93.3	97.4	91.7	77.6	91.1
ATTSD (Cai et al., 2025b)	ICCVW 2025	91.8	94.7	93.8	85.0	97.3	91.4	81.2	90.2	90.7
C2P-CLIP (Tan et al., 2025a)	AAAI2025	88.2	90.9	97.9	96.4	99.0	98.8	96.5	98.7	95.8
Ours	–	93.6	99.8	99.6	92.2	98.4	99.7	99.5	99.1	**97.7**

Cross-Semantic Evaluation. As shown in Table 3, we evaluate our cross-semantic generalization performance using two different training sources. When trained on UniverForen, which contains a limited set of only 20 semantic categories, most existing SOTA methods struggle to generalize to unseen concepts, yielding average accuracies below 82%. In contrast, our method proves to be highly semantic-agnostic and remains largely unaffected by this limited semantic diversity. When transitioning to the ImageNet-based training set, the generalization performance of most methods naturally improves due to the richer diversity of semantics in training data. While ATTSD specifically introduces a test-time semantic alignment strategy to tackle this exact scenario, our semantically unconstrained approach still outperforms such test-time adaptation methods without requiring extra inference overhead. Overall, our method consistently achieves SOTA performance across both training sources, demonstrating exceptionally robust cross-semantic generalization.

Cross-Dataset Evaluation. In this section, we reproduce several representative methods on the large-scale CF dataset. As shown in Table 4, the average accuracy of most existing methods fails to surpass 80%. Notably, by incorporating language-space priors and image captions as text input, C2P-CLIP demonstrates competitive performance. Nevertheless, its dependence on semantic concepts prevents it from overcoming the *Semantic Eclipsing* effect, which ultimately bottlenecks its cross-dataset generalization—a complex scenario that inherently encompasses cross-semantic shifts. In contrast, by optimizing desemanticized textual anchors, our approach successfully bypasses this limitation, yielding the highest overall cross-dataset generalization performance.

Table 3: Accuracy (ACC, %) comparison of **cross-semantic** generalization on the CSAIID dataset. Results are grouped based on different training sources. * indicates retrained models based on open-source code.

Method	Real Source	Forgery Model	Outdoor Scenes	People	Animals	Vehicles	Food	Complex	Avg. Acc.(%)
GLFF (Ju et al., 2023)	LSUN	ProGAN	63.03	61.68	55.88	63.93	61.58	62.00	61.35
NoDown (Gragnaniello et al., 2021)	LSUN	ProGAN	52.53	51.08	50.28	53.18	51.03	53.10	51.87
NoDown-S (Gragnaniello et al., 2021)	LSUN	StyleGAN	51.78	52.88	51.88	51.63	50.73	51.30	51.70
UnivFD (Ojha et al., 2023)	LSUN	ProGAN	64.28	64.73	62.38	72.19	61.73	63.35	64.78
DIRE (Wang et al., 2023)	LSUN	StyleGAN	50.23	49.87	49.67	50.08	50.03	49.45	49.89
AIDE (Yan et al., 2024)	LSUN	ProGAN	53.83	52.08	53.88	54.88	50.78	52.70	53.03
NPR (Tan et al., 2024b)	LSUN	ProGAN	71.64	73.54	66.68	71.09	70.89	74.15	71.33
ATTSD (Cai et al., 2025b)	LSUN	ProGAN	76.59	82.74	87.64	80.94	81.89	84.00	82.30
C2P-CLIP* (Tan et al., 2025a)	LSUN	ProGAN	80.99	79.84	82.99	85.04	77.54	80.25	81.11
Effort* (Yan et al., 2025)	LSUN	ProGAN	81.09	76.04	77.04	87.49	74.44	79.55	79.27
Ours	LSUN	ProGAN	91.55	89.54	89.04	94.60	87.64	89.80	**90.36**
DIRE (Wang et al., 2023)	ImageNet	ADM	50.03	49.52	48.82	49.57	49.52	49.15	49.44
AIDE (Yan et al., 2024)	ImageNet	SD1.4	71.04	69.93	74.79	72.79	77.34	74.45	73.38
NPR (Tan et al., 2024b)	ImageNet	SD1.4	86.19	79.79	80.74	87.09	86.04	84.30	84.02
DRCT (Chen et al., 2024)	ImageNet	SD1.4	63.42	63.78	64.48	63.63	58.68	64.05	62.85
ATTSD (Cai et al., 2025b)	ImageNet	SD1.4	86.14	90.75	92.90	90.70	91.25	90.75	90.42
C2P-CLIP* (Tan et al., 2025a)	ImageNet	SD1.4	97.55	98.35	96.90	98.30	94.30	97.10	97.08
Effort* (Yan et al., 2025)	ImageNet	SD1.4	92.40	90.35	91.75	93.95	94.35	92.00	92.46
Ours	ImageNet	SD1.4	98.95	99.00	99.15	99.15	99.35	99.15	**99.12**

Table 4: Accuracy (ACC, %) comparison of **cross-dataset** generalization from CF dataset to UniverForen (GAN-based subsets) and GenImage datasets. All methods are retrained on the CF dataset based on their open-source code.

	From CF to UniverForen							From CF to GenImage								Total
ACC	BigGAN	CycleGAN	GauGAN	ProGAN	StarGAN	StyleGAN	StyleGAN2	BigGAN	SDv1.4	VQDM	SDv1.5	Wukong	ADM	GLIDE	Midjourney	AVG
CF (Park & Owens, 2025)	63.08	81.83	74.15	82.37	**86.32**	87.28	87.22	63.37	64.08	62.66	63.85	63.97	62.17	62.97	61.16	71.10
NPR (Tan et al., 2024b)	79.45	60.90	71.13	73.54	95.30	67.43	68.31	74.57	80.04	77.88	80.42	79.73	79.98	79.70	77.93	76.42
UnivFD (Ojha et al., 2023)	60.55	63.21	74.35	52.81	79.31	53.45	55.25	77.22	76.78	75.42	77.26	75.64	74.89	76.28	74.82	69.82
C2P-CLIP (Tan et al., 2025a)	67.25	86.11	68.71	52.99	53.90	66.83	64.35	89.95	90.09	90.18	89.86	89.81	89.63	89.98	90.01	78.64
Ours	**84.25**	**89.48**	**84.77**	**82.71**	57.25	**89.08**	**88.92**	**92.23**	**92.55**	**92.39**	**92.50**	**92.28**	**91.32**	**92.04**	**92.15**	**87.59**

5.4 Ablation Study

To assess the effectiveness of our design choices, we construct five variants of our framework: (1) **CLIP**: Standard CLIP-style alignment using fixed hard prompts (e.g., "A photo of a real/fake image."). (2) **CoOP**: Conventional prompt tuning, which prepends learnable context tokens to explicit class labels ("real/fake"). (3) **CLIP-LoRA**: The purely visual adaptation shares the same LoRA setting as SUA, but without the contrastive learning branch. (4) **WoTextEnc**: A variant that removes the text encoder and directly contrasts visual features with two sets of learnable embedding vectors, bypassing the pretrained language space. (5) **AnchorInf**: The same training setting as SUA, but using anchor similarity of the contrastive head for inference instead of the binary classification head.

Table 5: Ablation study on the effectiveness of design choices, measured by accuracy. We compare **SUA** with task-adapted **CLIP** and **CoOP**, purely visual LoRA adaptation without contrastive learning (**CLIP-LoRA**), direct alignment with learnable tokens without a text encoder (**WoTextEnc**), and anchor-similarity inference under the same training configuration as SUA (**AnchorInf**).

Method	UniverForen	CSAIID	Avg.(%)
CLIP	87.79	81.19	84.49
CoOP	94.08	88.01	91.05
CLIP-LoRA	90.32	85.68	88.00
WoTextEnc	93.67	81.80	87.74
AnchorInf	85.33	79.23	82.08
Ours	**95.94**	**90.36**	**93.15**

As shown in Table 5, the standard **CLIP** variant yields poor performance, indicating that fixed real/fake descriptions are poorly equipped to reflect complex forgery traces. **CoOP**-style prompt tuning achieves higher accuracy, suggesting that incorporating learnable tokens helps adapt the latent space. The **WoTextEnc** variant suffers a significant performance drop compared to our full method, suggesting that bypassing the pretrained language space altogether harms generalizability. **CLIP-LoRA** also lags far behind, supporting the effectiveness of SUA's contrastive objective. The **AnchorInf** results suggest that contrastive alignment is most beneficial during training, allowing us to discard the contrastive branch during inference. Overall, our method achieves the best performance.

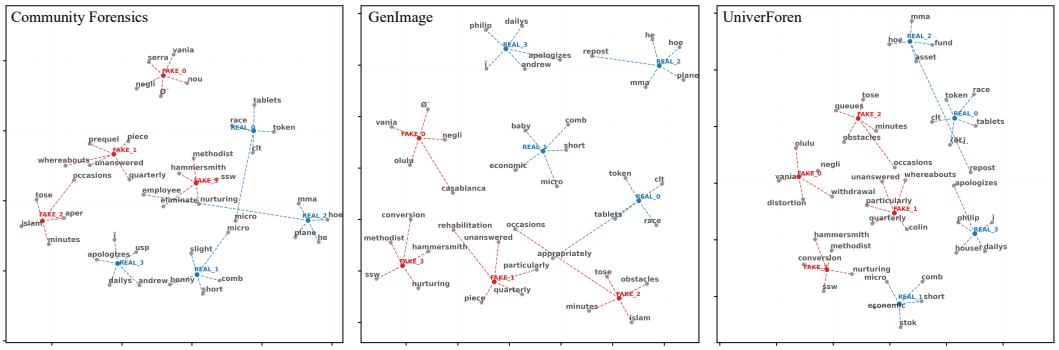

Figure 4: Latent space visualization of learned desemanticized textual anchors. We project the continuous tokens of real/fake anchors and their top-k discrete vocabulary neighbors within the CLIP latent space across three training datasets (CF, GenImage, and UniverForen). The learned `FAKE` (red) and `REAL` (blue) tokens avoid mapping to explicit categorical nouns or authenticity-related descriptors (e.g., *"real/fake"* labels or synonyms). Instead, most tokens anchor to sub-word fragments and meaningless typographical symbols. This cross-dataset visualization indicates that intrinsic forgery traces are largely decoupled from tangible, describable concepts.

5.5 Anchor Visualization

To better understand the behavior of the desemanticized textual anchors, we project the learned tokens of real/fake anchors and their top-k nearest discrete words in the CLIP vocabulary into a 2D space using t-SNE (Maaten & Hinton, 2008). We extract and compare the tokens trained independently across three datasets (CF, GenImage, and UniverForen). This visualization allows us to investigate how our SUA-based representations behave under varying training semantic distributions.

As illustrated in Fig. 4, none of the learned tokens project closely to explicit categorical nouns, nor do they exhibit clear semantic alignment with explicit category or authenticity words (e.g., "*real/fake*" labels or their synonyms). Instead, their nearest neighbors consist predominantly of abstract vocabulary, sub-word fragments (e.g., "*mma*", "*ssw*"), and even meaningless typographical symbols. This phenomenon aligns perfectly with our core motivation for mitigating Semantic Eclipsing: it suggests that true forgery artifacts are not tied to tangible semantic concepts. Furthermore, we observe remarkable cross-dataset stability. Even when trained on entirely different semantic distributions, the tokens exhibit consistent spatial anchoring. For instance, across all three datasets, the `REAL_0` token is consistently associated with the same neighboring sub-words like "*clt*", "*race*", and "*token*". Similarly, the `FAKE_2` token shows consistent associations with words such as "*tose*", "*minutes*", and "*occasions*". Interestingly, rather than all tokens behaving uniformly, we observe a clear division of labor: While some tokens adapt to dataset-specific variations, those consistently observed across datasets suggest that our method provides generalizable anchors in the language space for forgery detection.

6 Conclusion

In this paper, we identify a critical vulnerability in current VLM-based forensic detectors—*Semantic Eclipsing*—where the strong semantic inductive bias in pretrained representations overshadows generation artifacts. To mitigate this effect, we propose a semantically unconstrained alignment strategy. By aligning visual features with continuous learnable tokens, termed desemanticized textual anchors, our method successfully maps intrinsic forgery traces to desemanticized regions of the latent space. Extensive empirical evaluations demonstrate that our detector achieves state-of-the-art generalization across diverse cross-generator, cross-dataset, and cross-semantic benchmarks. Ultimately, our findings challenge the necessity of explicit natural-language inputs in CLIP-adapted detectors, paving the way for more resilient, domain-agnostic AIGI detection.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for writing assistance. We have not used generative AI tools for idea generation or related work retrieval, and formulating mathematical claims, or writing mathematical proofs; are not applicable to this work. Additionally, we used generative AI tools for code-writing for data analysis. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details. This statement should not be more than 1 page.

ETHICS STATEMENT

This work focuses on improving the generalization of AI-generated image detection. The proposed method is intended to support media authenticity and forensic analysis. As with other detection methods, the insights from this work may potentially be used to study or circumvent existing detectors. We therefore encourage responsible use of the proposed method and evaluation results for research and defensive applications.

REPRODUCIBILITY STATEMENT

We provide detailed implementation settings to facilitate the reproducibility of our experiments. The model selection, GPU configuration, experimental framework, training protocol, and other implementation details are described in Sec. 5.1. The selection and validation of key hyperparameters are reported through ablation studies in Sec. A.3. Together, these details specify the experimental setup necessary to reproduce the main results reported in this work.

REFERENCES

Yu Cai, Shan Jia, Jiahe Tian, Jiao Dai, Jizhong Han, and Siwei Lyu. Cataid: Category-guided ai-generated image detection via vision-language model adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a.

Yu Cai, Jiahe Tian, Xiaomeng Fu, Jiao Dai, Jizhong Han, and Siwei Lyu. Adaptive test-time semantic debiasing for ai-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *ECCV*, 2020.

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 *Forty-first International Conference on Machine Learning*, 2024.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

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

Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. In *2021 IEEE international conference on multimedia and expo (ICME)*, 2021.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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

Yonghyun Jeong, Doyeon Kim, Seungjai Min, Seongho Joe, Youngjune Gwon, and Jongwon Choi. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 48–57, 2022.

Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, 2022.

Yan Ju, Shan Jia, Jialing Cai, Haiying Guan, and Siwei Lyu. Glff: Global and local feature fusion for ai-synthesized image detection. *IEEE Transactions on Multimedia*, 2023.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution ai-generated image detection by spectral learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18706–18717. IEEE, 2025.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision (ECCV)*. Springer, 2024.

Francesco Laiti, Benedetta Liberatori, Thomas De Min, and Elisa Ricci. Conditioned prompt-optimization for continual deepfake detection. In *International Conference on Pattern Recognition*, 2024.

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

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008.

Lakshmanan Nataraj, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, Amit K Roy-Chowdhury, and BS Manjunath. Detecting gan generated fake images using co-occurrence matrices. *arXiv preprint arXiv:1903.06836*, 2019.

Joao C Neves, Ruben Tolosana, Ruben Vera-Rodriguez, Vasco Lopes, Hugo Proença, and Julian Fierrez. Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1038–1048, 2020.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

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

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*, 2020.

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

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, 2023.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

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

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

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

Chuangchuang Tan, Jinglu Wang, Xiang Ming, Renshuai Tao, Yunchao Wei, Yao Zhao, and Yan Lu. Forenx: Towards explainable ai-generated image detection with multimodal large language models. *arXiv preprint arXiv:2508.01402*, 2025b.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 2021.

Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. Stylegan2 distillation for feed-forward image manipulation. In *European conference on computer vision*, 2020.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, 2020.

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

Siwei Wen, Junyan Ye, 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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Haiwei Wu, Jiantao Zhou, and Shile Zhang. Generalizable synthetic image detection via language-guided contrastive learning. *IEEE Transactions on Artificial Intelligence*, 2025.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. In *International Conference on Machine Learning*, 2025.

Muli Yang, Gabriel James Goenawan, Henan Wang, Huaiyuan Qin, Chenghao Xu, Yanhua Yang, Fen Fang, Ying Sun, Joo Hwee Lim, and Hongyuan Zhu. Your ai-generated image detector can secretly achieve sota accuracy, if calibrated. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11622–11630, 2026.

Cai Yu, Peng Chen, Jiao Dai, Xi Wang, Weibo Zhang, Jin Liu, and Jizhong Han. Focus by prior: Deepfake detection based on prior-attention. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, 2022.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11941–11952. IEEE, 2023.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, 2019.

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.

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, 2024.

A APPENDIX

A.1 ANALYSIS OF CLIP-ADAPTED DETECTORS

Naive CLIP-LoRA Detector. We first train a baseline detector by applying LoRA fine-tuning to the visual encoder of a pretrained CLIP backbone on the UniverForen training set, thereby adapting CLIP's general-purpose visual space towards a forensic space. We refer to this model as the CLIP-LoRA detector.

Classification behavior: We then analyse its behavior on four subsets of the UniverForen test set — StyleGAN, StyleGAN2, Guided Diffusion, and Deepfake — by decomposing every prediction into the four confusion-matrix outcomes. Taking fake as the positive class, these are: true positives (TP), fake images correctly predicted as fake; true negatives (TN), real images correctly predicted as real; false positives (FP), real images incorrectly predicted as fake; and false negatives (FN), fake images incorrectly predicted as real. Fig. 5 reports the four groups arranged by predicted label, i.e., TN and FN together form the "predicted real" side and TP and FP the "predicted fake" side, with the corresponding per-group rate (TNR, FNR, TPR, FPR) printed beneath each group.

Feature-space probe: For each test image, we additionally measure how far the forensics-adapted representation has drifted from the original one. Concretely, we compute the cosine similarity between the image embedding produced by the LoRA-adapted visual encoder and the embedding produced by the frozen, original CLIP visual encoder. This quantity forms the vertical axis of the figure and acts as a per-image measure of how much of the original CLIP visual feature survives forensic adaptation.

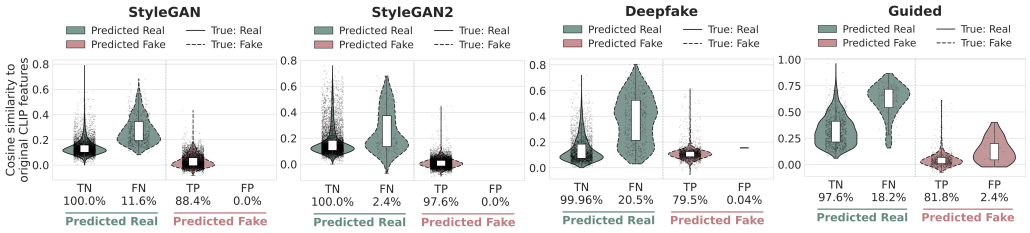

Figure 5: *Authenticity-oriented* discrimination tendency of the CLIP-LoRA detector: higher similarity to original CLIP features is associated with a tendency to predict images as real. We present per-image cosine similarity between the LoRA-adapted and the original (frozen) CLIP visual features, for the four confusion-matrix outcomes of the CLIP-LoRA detector on four UniverForen test subsets. Fake is the positive class. Percentages below each group are TNR, FNR, TPR and FPR, respectively.

Observation: Across all four subsets, the CLIP-LoRA detector tends to predict images whose adapted features remain more similar to their original CLIP features as real. Notably, the FN distribution is centered at a markedly higher cosine similarity than either the TN or TP distribution, indicating that missed fake images retain particularly strong feature alignment with their original representations. This pattern suggests an *authenticity-oriented* discrimination tendency: stronger retention of the original CLIP features is associated with real-image predictions, even for unseen fakes. This observation is consistent with our motivation that insufficient adaptation beyond CLIP's pretrained semantic inductive bias may hinder fake-image identification.

Forensics-Specific CLIP-Adapted Detectors. To better illustrate the default-to-real tendency of CLIP-adapted models, we examine two representative forensics-specific methods based on CLIP, DeFake (Sha et al., 2023) and C2P-CLIP (Tan et al., 2025a). As shown in Fig. 6, our analysis reveals a persistent propensity to predict unseen fake images as real. A related observation was made by UnivFD (Ojha et al., 2023) regarding CNN-based detectors, demonstrating that the real class collapses into a "sink class" that hosts anything that does not resemble the training-time GAN fakes. In our analysis, we extend this investigation to CLIP-based detectors through a controlled generalization study under a **semantic-alignment** training regime. Specifically, we train DeFake and C2P-CLIP using only one semantic category (Real Horse vs. ProGAN (Karras et al., 2018) Horse) from the UniverForen dataset. This setup ensures that semantic categories are highly aligned during training, encouraging the model to isolate GAN-specific artifacts.

An ideal detector should reliably identify both real and fake images. However, when evaluated on unseen StyleGAN2 (Viazovetskyi et al., 2020) images, both detectors exhibit a consistent *default-to-real* tendency, misclassifying the majority of fake samples as real. This failure highlights a critical limitation: these detectors are not learning universal forensic traces; instead, they associate fake clues with the specific semantic-generator correlations seen during training. Once an unfamiliar fake sample deviates from these training-time correlations, it is overwhelmingly absorbed by the massive real manifold embedded in CLIP's pre-training, yielding a default "real" prediction.

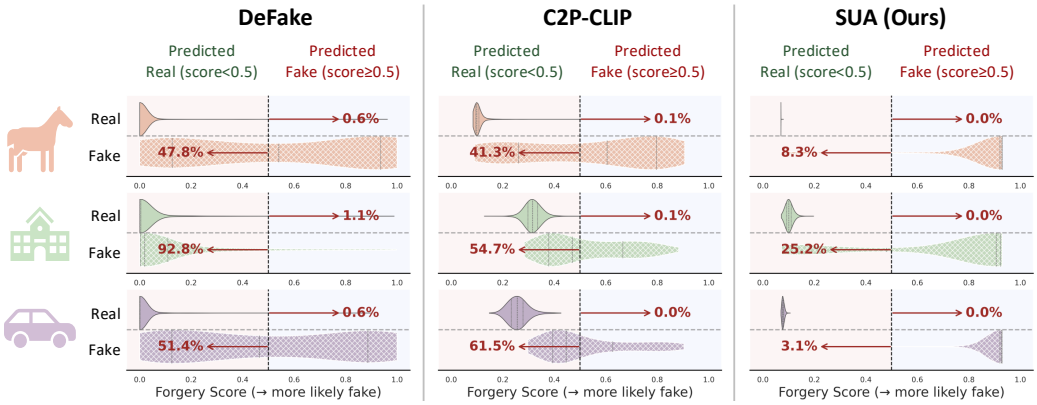

Figure 6: Supporting Evidence of *default-to-real* preference. Models are trained under the **semantic-alignment** setting using the horse-only images (Real vs. ProGAN) and evaluated on unseen images across different semantic categories (Real vs. StyleGAN2). Red values indicate misclassification rates. While baseline detectors (DeFake (Sha et al., 2023), C2P-CLIP (Tan et al., 2025a)) overwhelmingly misclassify unseen fakes as real, our SUA preserves distinct real/fake separability. Best viewed in color.

A.2 SUA AS A PLUG-AND-PLAY ENHANCEMENT ON EXISTING BASELINES

SUA is straightforward to implement and can be integrated into existing CLIP-based detectors as an additional enhancement. To evaluate its compatibility with different adaptation strategies, we consider CLIP-LoRA, a baseline using LoRA-based visual adaptation, and Effort, a forgery-specific detector employing SVD-based fine-tuning. Both approaches adapt CLIP's visual backbone on forgery data through parameter-efficient tuning. We augment each baseline with SUA's learnable textual anchors and image–anchor contrastive objective, while retaining its original visual adaptation strategy.

As shown in Table 6, SUA consistently improves both ACC and AP across all six semantic categories. For CLIP-LoRA, average ACC and AP increase by 4.68 and 3.89 percentage points, respectively; for Effort, the corresponding gains are 3.56 and 2.91 percentage points. Importantly, the AP improvements indicate better ranking of real and fake images beyond gains at a fixed classification threshold, supporting the learning of more discriminative representations. These results demonstrate that SUA complements both generic and forgery-specific CLIP adaptation strategies, providing a simple and effective enhancement for existing methods.

To study how SUA transfers to other vision–language foundation models, we additionally adopt SigLIP (Zhai et al., 2023) as a backbone. Specifically, we choose SigLIP-Large/16, whose visual tower is matched in capacity to CLIP ViT-L/14 (316M vs. 304M parameters), so the visual encoders have a comparable parameter count. We then evaluate SigLIP-LoRA with and without SUA. Although SigLIP-LoRA is markedly weaker than its CLIP counterpart in absolute terms, adding SUA still improves both ACC and AP performance. The lower performance may stem from directly transferring CLIP-tuned hyperparameters to SigLIP without further optimization. Nevertheless, the consistent gain indicates that SUA is not tied to a single backbone and can be plugged into other VLM-based detectors.

Table 6: **SUA as a plug-and-play mechanism for existing methods.** We evaluate cross-semantic generalization from UniverForen to CSAIID dataset using **ACC** and **AP**. Shaded rows augment the corresponding baseline with SUA, improving both average metrics across all six semantic categories. Δ denotes the gain in average performance over the baseline, measured in percentage points.

Method	Outdoor Scenes	People	Animals	Vehicles	Food	Complex	Avg. Acc.(%)	Δ
CLIP-LoRA	86.59	86.84	85.19	91.35	79.49	84.60	85.68	–
CLIP-LoRA+SUA	91.55	89.54	89.04	94.60	87.64	89.80	90.36	**+4.68**
Effort	81.09	76.04	77.04	87.49	74.44	79.55	79.27	–
Effort+SUA	84.24	78.74	83.04	90.90	77.64	82.40	82.83	**+3.56**
SigLIP-LoRA	73.78	71.83	71.08	76.88	71.23	72.85	72.94	–
SigLIP-LoRA+SUA	79.18	77.28	75.28	71.33	74.73	78.65	76.08	**+3.14**
Method	**Outdoor Scenes**	**People**	**Animals**	**Vehicles**	**Food**	**Complex**	**Avg. AP.(%)**	**Δ**
CLIP-LoRA	93.08	93.40	92.04	95.96	88.17	91.85	92.42	–
CLIP-LoRA+SUA	97.02	96.55	95.44	98.49	93.93	96.42	96.31	**+3.89**
Effort	93.85	92.81	93.92	96.96	90.14	92.86	93.42	–
Effort+SUA	96.75	95.36	96.89	98.48	94.29	96.23	96.33	**+2.91**
SigLIP-LoRA	89.35	86.54	84.41	90.61	81.44	86.13	86.41	–
SigLIP-LoRA+SUA	88.88	88.50	87.98	90.91	82.11	87.33	87.62	**+1.21**

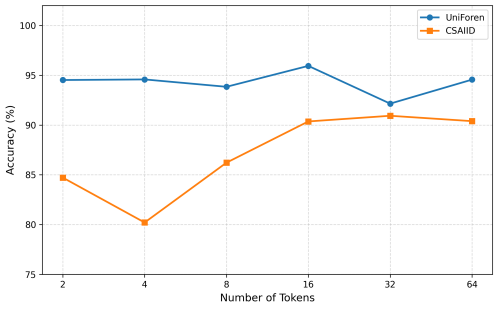

Figure 7: Effect of token number M on average detection accuracy across UniverForen and CSAIID subsets.

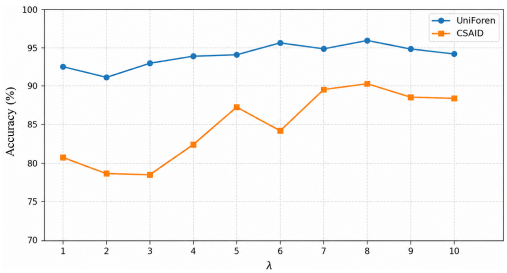

Figure 8: Impact of the loss weight λ on average detection accuracy across UniverForen and CSAIID subsets.

A.3 Hyperparameter Ablation

We investigate two hyperparameters in SUA: the number of learnable tokens per class (M) and the classification loss weight (λ). All variants are trained on UniverForen and evaluated on both UniverForen and CSAIID, reporting average detection accuracy across all subsets. We vary one hyperparameter at a time while keeping the other fixed at its default value.

Number of Learnable Tokens. As illustrated in Fig. 7, increasing the token capacity from 2 to 16 initially yields significant performance gains. This general trend suggests that a larger number of tokens provides stronger representational capacity for capturing the diverse, multifaceted generative artifacts. However, the performance gains gradually saturate beyond 16 tokens. This indicates that while additional tokens enhance the representational capability of the desemanticized anchors, an excessive number of tokens leads to diminishing returns. We therefore adopt $M = 16$ as the default, balancing performance across the two benchmarks with a compact token representation.

Classification Loss Weight. The coefficient λ balances the contribution of the classification and contrastive losses in the joint objective $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{con}} + \lambda\mathcal{L}_{\text{cls}}$. A smaller λ places greater relative emphasis on image–anchor contrastive alignment, whereas a larger λ increases the contribution of binary classification supervision. We vary λ to examine how this balance affects detection

Table 7: **Cross-resolution** evaluation on the SPAI benchmark. Models are trained on 180K latent-diffusion images paired with 180K real images from COCO and LSUN. Results report **Balanced Accuracy (ACC %)** at a threshold of 0.5, averaged over five real-image sources for each generator.

Approach	Image Size < 0.5 MPixels			Image Size 0.5–1.0 MPixels						Image Size > 1.0 MPixels				Acc.
	Glide	SD1.3	SD1.4	Flux	DALLE2	SD2	SDXL	SD3	GigaGAN	MJv5	MJv6.1	DALLE3	Firefly	AVG
DIRE (Wang et al., 2023)	38.2	57.7	58.5	46.0	51.5	61.7	47.3	49.6	40.9	43.8	50.4	62.6	51.1	50.7
CNNDet. (Wang et al., 2020)	52.0	52.2	52.3	49.1	58.0	52.2	56.6	47.5	63.0	50.8	52.6	47.6	55.5	53.0
NPR (Tan et al., 2024b)	79.9	79.9	79.4	29.9	29.9	30.1	29.9	79.9	79.9	31.4	29.9	79.9	29.9	53.1
Fusing (Ju et al., 2022)	54.1	52.9	53.0	50.5	59.4	52.6	51.4	48.2	62.6	53.2	58.6	48.2	55.9	53.9
FreqDet. (Frank et al., 2020)	46.9	80.4	80.6	42.6	42.8	46.3	61.8	65.1	60.9	44.0	41.6	46.9	77.2	56.7
UnivFD (Ojha et al., 2023)	50.8	67.9	67.4	45.9	81.0	73.2	65.5	45.0	73.6	50.0	53.8	44.9	88.2	62.1
LGrad (Tan et al., 2023)	69.3	78.6	78.8	68.9	79.2	54.6	64.3	34.0	80.4	63.3	73.6	37.2	42.1	63.4
GramNet (Liu et al., 2020)	72.7	78.7	79.1	73.9	80.4	58.5	72.3	32.4	80.4	58.2	80.3	37.0	34.8	64.5
DeFake (Sha et al., 2023)	76.3	59.5	59.0	79.6	45.2	61.7	51.7	78.3	67.0	62.2	77.1	81.5	44.0	64.8
DMID (Corvi et al., 2023)	52.4	99.3	99.3	82.1	49.5	97.9	96.5	57.9	54.3	98.4	78.1	49.3	57.5	74.8
PatchCr. (Zhong et al., 2023)	69.6	86.6	86.7	79.7	71.7	86.8	89.4	40.2	89.0	72.6	87.9	40.6	72.5	74.9
RINE (Koutlis & Papadopoulos, 2024)	88.5	96.6	96.4	84.1	82.0	89.6	95.5	47.0	83.3	90.1	69.2	47.6	67.0	79.8
SPAI (Karageorgiou et al., 2025)	81.3	92.2	92.3	71.9	83.4	88.8	90.5	61.3	76.2	87.2	74.3	80.8	90.3	82.3
Ours	84.5	96.4	96.4	58.5	93.3	93.1	96.3	48.8	90.4	94.3	79.6	48.9	93.7	**82.6**

Table 8: Analogous results on **cross-resolution** generalization, where we use **Average Precision (AP, %)** to compare the methods.

Approach	Image Size < 0.5 MPixels			Image Size 0.5–1.0 MPixels						Image Size > 1.0 MPixels				AP
	Glide	SD1.3	SD1.4	Flux	DALLE2	SD2	SDXL	SD3	GigaGAN	MJv5	MJv6.1	DALLE3	Firefly	AVG
NPR (Tan et al., 2024b)	75.4	87.6	72.3	39.3	31.3	35.2	38.3	72.4	84.0	38.2	31.0	97.4	50.4	57.9
DIRE (Wang et al., 2023)	39.5	56.6	57.4	45.2	52.2	70.2	47.8	48.0	41.7	44.1	40.1	59.7	48.6	50.1
CNNDet. (Wang et al., 2020)	58.0	57.6	59.4	44.1	70.8	57.2	67.0	37.8	75.6	50.6	47.2	35.5	69.2	56.1
FreqDet. (Frank et al., 2020)	46.3	90.6	90.8	42.6	45.2	46.7	60.5	64.7	63.7	43.6	30.5	47.5	71.7	57.3
Fusing (Ju et al., 2022)	64.7	63.6	63.0	57.7	77.1	65.6	61.9	42.1	80.7	64.6	67.6	36.6	74.0	63.0
LGrad (Tan et al., 2023)	80.1	76.8	78.4	71.7	83.1	59.3	69.2	32.6	89.7	67.2	69.3	38.3	44.6	66.2
UnivFD (Ojha et al., 2023)	62.6	82.1	81.9	43.3	91.4	86.1	79.9	38.6	87.0	57.9	52.9	39.6	95.9	69.2
GramNet (Liu et al., 2020)	74.9	79.4	79.8	74.6	80.3	65.2	74.0	36.2	80.1	63.5	73.3	48.0	45.5	67.3
DeFake (Sha et al., 2023)	87.0	63.0	62.5	90.8	44.9	66.7	54.7	87.4	73.9	66.6	83.8	93.2	42.9	70.6
PatchCr. (Zhong et al., 2023)	79.7	96.3	96.7	86.8	80.0	96.1	95.3	41.5	97.7	81.1	94.1	39.0	77.5	81.7
DMID (Corvi et al., 2023)	71.5	100.0	100.0	97.2	54.9	99.7	99.7	71.8	70.2	99.9	93.2	45.3	87.7	83.9
RINE (Koutlis & Papadopoulos, 2024)	95.5	99.9	99.9	94.0	93.2	96.8	99.4	46.9	93.4	97.0	78.0	48.6	83.1	86.6
SPAI (Karageorgiou et al., 2025)	90.9	99.3	99.4	82.0	90.7	96.8	97.2	72.4	84.7	95.1	79.1	88.8	94.9	90.1
Ours	94.1	100.0	100.0	80.7	98.1	98.7	99.9	52.2	97.4	98.5	91.0	63.3	98.0	**90.2**

performance, as shown in Fig. 8. Increasing λ generally improves accuracy on CSAIID up to $\lambda = 8$, despite some fluctuations, while performance on UniverForen remains relatively stable. Both benchmarks achieve their highest accuracy at $\lambda = 8$, followed by a decline at larger λ values. We therefore set $\lambda = 8$ as the default to balance classification supervision and contrastive alignment.

A.4 Cross-Resolution Evaluation

Following SPAI (Karageorgiou et al., 2025), we train our models on the 180K latent-diffusion images of (Corvi et al., 2023), paired with the 180K real images drawn from COCO and LSUN, and evaluate on the SPAI test set. This setting is inherently cross-resolution: all generated training images are 256×256, whereas the test set spans three resolution ranges, from below 0.5 to above 1.0 megapixels. We strictly follow their test protocol: We adopt SPAI's patch-based inference setting: the image-level score is the mean of the patch-level predictions; Each metric is computed per generator against each source of real images and then averaged over the five sources. Balanced accuracy and average precision are reported in Table 7 and Table 8, respectively.

Our method achieves the highest average performance among the compared methods on both metrics, reaching 82.6% balanced accuracy and 90.2% average precision. Although the improvements over SPAI are modest, it is worth noting that SPAI employs strong data augmentation and four image views for spectral learning, whereas our semantic-eclipsing suppression strategy is not specifically designed for this cross-resolution setting. These results highlight the potential of CLIP-based detectors for cross-resolution generalization.

A.5 Additional Evaluation Metrics: Average Precision

In the main manuscript (Section 5), we report the detection Accuracy to evaluate the generalization performance of our proposed Semantically Unconstrained Alignment (SUA) framework. To provide a more comprehensive assessment of our detector, this section presents the Average Precision (AP) results for both the cross-generator and cross-semantic generalization benchmarks.

Table 9: **Average Precision (AP, %)** comparison of **cross-generator** generalization on the UniverForen dataset. All methods were trained on the ProGAN-generated images vs. real images from LSUN and evaluated on subsets from different generators.

Method	Generative Adversarial Networks						Deep fakes	Low level vision		Perceptual loss		Guided	LDM			Glide			DA LLE	Total
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/ CFG	100 steps	100 27	50 27	100 10		Avg. AP(%)
CNNSpot (Wang et al., 2020)	100.0	93.47	84.50	99.54	89.49	98.15	89.02	73.75	59.47	98.24	98.40	73.72	70.62	71.00	70.54	80.65	84.91	82.07	70.59	83.58
Patchfor (Chai et al., 2020)	80.88	72.84	71.66	85.75	65.99	69.25	76.55	76.19	76.34	74.52	68.52	75.03	87.10	86.72	86.40	85.37	83.73	78.38	75.67	77.73
Co-occurence (Nataraj et al., 2019)	99.74	80.95	50.61	98.63	53.11	67.99	59.14	68.98	60.42	73.06	87.21	70.20	91.21	89.02	92.39	89.32	88.35	82.79	80.96	78.11
Spec (Zhang et al., 2019)	55.39	100.0	75.08	55.11	66.08	100.0	45.18	47.46	57.12	53.61	50.98	57.72	77.72	77.25	76.47	68.58	64.58	61.92	67.77	66.21
UnivFD (Ojha et al., 2023)	100.0	98.13	94.46	86.66	99.25	99.53	91.67	78.54	67.54	83.12	91.06	79.24	95.81	79.77	95.93	93.93	95.12	94.59	88.45	90.14
NPR* (Tan et al., 2024b)	100.0	99.53	94.53	99.94	88.82	100.0	84.41	97.95	99.99	50.16	50.16	98.26	99.92	99.91	99.92	99.87	99.89	99.92	99.26	92.76
LGrad(Tan et al., 2023)	100.0	93.98	90.69	99.86	79.36	99.98	67.91	59.42	51.42	63.52	69.61	87.06	99.03	99.16	99.18	93.23	95.10	94.93	97.23	86.35
FreqNet (Tan et al., 2024a)	99.92	99.63	96.05	99.89	99.71	98.63	99.92	94.42	74.59	80.10	75.70	96.27	96.06	100.0	62.34	99.80	99.78	96.39	77.78	91.95
FatFormer (Liu et al., 2024)	100.0	100.0	99.98	99.75	100.0	100.0	97.99	97.94	81.21	99.84	99.93	91.99	99.81	99.09	99.87	99.13	99.41	99.20	99.82	98.16
ATTSD (Cai et al., 2025b)	100.0	98.39	94.80	85.27	99.38	100.00	58.49	69.44	61.21	99.24	99.48	81.12	99.93	99.77	99.93	99.82	99.88	99.84	99.93	91.89
C2P-CLIP* (Tan et al., 2025a)	100.0	100.0	99.94	99.49	100.0	100.00	98.82	96.51	98.90	99.89	99.97	92.04	99.98	99.75	99.98	99.79	99.92	99.93	99.83	99.20
Effort (Yan et al., 2025)	100.0	100.0	99.99	99.77	100.0	100.0	98.95	97.53	97.53	100.0	100.0	95.39	99.99	99.89	100.0	99.87	99.92	99.98	99.96	99.41
Ours	100.0	100.0	99.99	99.95	100.0	100.0	98.99	97.65	99.93	99.96	94.29	99.99	99.97	100.0	99.94	99.94	99.97	99.99	98.65	99.43

Cross-Generator Generalization (AP). Table 9 details the AP performance of various methods on the UniverForen benchmark. All models are trained exclusively on ProGAN-generated images paired with LSUN real images and evaluated across 19 unseen generative models. Consistent with the accuracy metrics reported in the main text, our SUA method achieves the best average AP of 99.43%. Notably, it maintains exceptionally high precision across nearly all unseen domains. This confirms that our desemanticized anchors reliably capture generator-agnostic forgery traces and maintain high confidence in their predictions.

Cross-Semantic Generalization (AP). Table 10 presents the analogous AP results on the CSAIID benchmark to evaluate cross-semantic generalization. Same as in the main text, we report results under two distinct training settings: (1) LSUN & ProGAN and (2) ImageNet & Stable Diffusion v1.4. Under the highly restricted LSUN & ProGAN setting, where the training semantic diversity is limited, our method still achieves an impressive average AP of 96.31%, outperforming competitive VLM-based detectors such as ATTSD and C2P-CLIP. Furthermore, when trained on the more diverse ImageNet & SD1.4 dataset, our method achieves a near-perfect average AP of 99.99%. These AP metrics strongly reinforce the core premise of our work: Guided by desemanticized anchors, our framework effectively decouples microscopic forgery traces from semantic variations, ensuring highly generalizable detection across entirely unseen semantic categories.

Table 10: Analogous results of **cross-semantic generalization** on CSAIID, where we use **Average Precision (AP, %)** to compare the methods.

Method	**Real Source**	**Forgery Model**	**Outdoor Scenes**	**People**	**Animals**	**Vehicles**	**Food**	**Complex**	**Avg. AP.(%)**
GLFF (Ju et al., 2023)	LSUN	ProGAN	38.21	40.23	47.99	38.65	39.76	38.88	40.62
NoDown (Gragnaniello et al., 2021)	LSUN	ProGAN	74.15	63.02	56.10	77.66	65.83	73.55	68.39
NoDown-S (Gragnaniello et al., 2021)	LSUN	StyleGAN	76.09	72.51	75.98	76.06	74.56	74.18	74.90
UnivFD (Ojha et al., 2023)	LSUN	ProGAN	76.21	79.08	75.59	78.09	76.03	75.45	76.74
DIRE (Wang et al., 2023)	LSUN	StyleGAN	47.43	45.46	41.63	43.00	45.15	42.99	44.28
AIDE (Yan et al., 2024)	LSUN	ProGAN	77.78	78.90	78.28	80.26	66.50	76.87	76.43
NPR (Tan et al., 2024b)	LSUN	ProGAN	81.71	82.56	78.29	80.01	81.37	72.92	79.48
ATTSD (Cai et al., 2025b)	LSUN	ProGAN	87.89	93.96	95.34	93.00	90.01	93.50	92.28
C2P-CLIP* (Tan et al., 2025a)	LSUN	ProGAN	96.19	95.15	96.62	96.73	94.43	96.24	95.89
Effort* (Yan et al., 2025)	LSUN	ProGAN	93.85	92.81	93.92	96.96	90.14	92.86	93.42
Ours	LSUN	ProGAN	97.02	96.55	95.44	98.49	93.93	96.42	96.31
DIRE (Wang et al., 2023)	ImageNet	ADM	50.73	42.71	41.87	49.53	38.05	41.66	44.09
AIDE (Yan et al., 2024)	ImageNet	SD1.4	94.54	95.53	96.07	95.48	96.26	94.89	95.46
NPR (Tan et al., 2024b)	ImageNet	SD1.4	94.64	90.93	88.73	96.04	94.10	94.05	93.08
DRCT (Chen et al., 2024)	ImageNet	SD1.4	57.95	58.14	57.91	58.02	54.80	58.36	57.53
ATTSD (Cai et al., 2025b)	ImageNet	SD1.4	95.45	97.76	98.54	97.79	97.21	97.48	97.37
C2P-CLIP* (Tan et al., 2025a)	ImageNet	SD1.4	99.56	99.79	99.80	99.88	99.79	99.70	99.76
Effort* (Yan et al., 2025)	ImageNet	SD1.4	99.08	98.84	98.86	99.42	99.31	99.37	99.15
Ours	ImageNet	SD1.4	99.96	99.99	99.99	100.00	100.00	99.99	99.99

A.6 Limitations and Future Work

Our method is built upon the pretrained CLIP text encoder, which provides a rich, implicit semantic prior that continuous learnable tokens can repurpose for forensic clues. The main assumption underlying our method is that the strong semantic inductive bias inherited from CLIP's pretraining limits its effectiveness in artifact capturing. However, we do not rule out the possibility that a large-scale vision-language model trained exclusively on massive and highly diverse datasets—composed entirely of forged visual content paired with artifact-related textual descriptions—could better leverage the language utility for forensic understanding. Although our anchor visualizations can provide some spatial interpretability, the deliberate decoupling of our anchors from human-readable vocabulary limits post-hoc linguistic interpretability. In future work, we plan to explore combining semantically unconstrained cross-modal alignment with lightweight decoding schemes (e.g., concept decomposition or nearest sub-word attribution over a constrained vocabulary), aiming to interpret what each learned token encodes. We expect such approaches to enhance the interpretability of forgery representations without re-introducing semantic shortcuts during training, thereby providing better support for post-hoc forensic analysis.

A.7 Broader Impact

This work improves the generalization of AI-generated image detection by mitigating semantic bias in vision-language models, enabling more reliable detection across unseen generators and domains, which is important for combating misinformation and deepfake misuse. However, the proposed insights into generator-specific artifacts may also introduce dual-use risks, as they could potentially guide the development of more advanced generation models that evade detection.