Too Aligned to be Real: Detecting AI-Generated Images via Cross-modal Alignment Shift

Haifeng Zhang[1, 3], **Qinghui He**[1, 3], **Xiuli Bi**[2, 3], **Bo Liu**[2, 3,*], **Chi-Man Pun**[4], **Bin Xiao**[2, 3, 5,*]
[1]School of Computer Science and Technology, Chongqing University of Posts and Telecommunications
[2]School of Artificial Intelligence, Chongqing University of Posts and Telecommunications
[3]Chongqing Key Laboratory of Image Cognition, Chongqing, China
[4]Faculty of Information Science and Computing, University of Macau, Macao, China
[5]Jinan Inspur Data Technology Co., Ltd., Jinan, China
{D240201065, D250201011}@stu.cqupt.edu.cn, {bixl, boliu, xiaobin}@cqupt.edu.cn
cmpun@umac.mo

Abstract

AI-generated image detection has become increasingly important as generative models produce highly realistic visual content. Existing detectors mainly rely on visual artifacts or visual features extracted from pretrained multimodal models, while largely overlooking the structural differences between real and generated images in the joint image–text space. In this paper, we reveal a *Cross-modal Alignment Shift* (CAS): generated images tend to exhibit stronger and more concentrated alignment with text than semantically matched real images. We verify this phenomenon through retrieval preference, image–text residual structure, spectral concentration, and matching entropy. Motivated by this finding, we propose a CAS-based detection framework (CASNet) that learns an amplified multimodal model with a text-guided alignment-shift objective. The amplification-induced feature increment is extracted as an explicit cross-modal cue and coupled with amplified visual representations, enabling the detector to exploit both alignment-level structural evidence and visual discriminative cues. Extensive experiments across diffusion models, GANs, autoregressive models, and commercial generators demonstrate consistent improvements over state-of-the-art detectors. In particular, on recent commercial generators, our method improves *ACC* and *AP* by 5.48% and 6.13%, respectively. Code is available at `https://github.com/oceanzhf/CASNet`.

1 Introduction

The rapid advancement of generative technologies has enabled GANs [12, 19], diffusion models [35, 16], and their variants to produce highly photorealistic images that are increasingly present in daily life [47]. However, their misuse may spread misinformation and undermine trust in digital media [53, 23, 24, 25, 26], highlighting the urgent need for reliable AI-generated image (AIGI) detection methods.

Existing AIGI detectors mainly rely on visual cues such as spatial artifacts, frequency patterns, noise inconsistencies, and texture anomalies [46, 57, 11, 29, 28, 27, 56]. While effective for specific generators, they often generalize poorly to unseen models or high-quality images with subtle artifacts. Recent methods improve generalization by using pretrained multimodal foundation models as universal feature extractors [31, 54, 51, 55], but they mainly consume the visual branch and do not measure the structural difference between real and generated images in the joint image–text space itself.

*Corresponding Authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

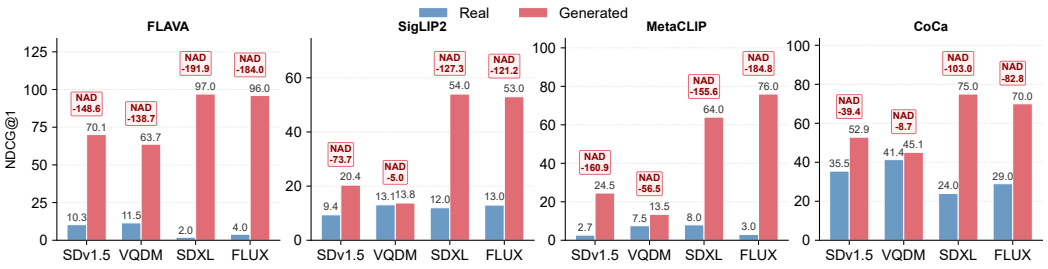


Figure 1: Cross-modal alignment shift in paired image–text data. With four pretrained vision–language backbones (FLAVA [38], SigLIP2 [44], MetaCLIP [49], and CoCa [52]), generated images obtain higher NDCG@1 than matched real images across generators. More negative NAD values, computed from the real and generated NDCG@1 scores, indicate a stronger alignment discrepancy in favor of generated images.

Meanwhile, several works have attempted to incorporate text into AIGI detection by formulating the task as prompt learning, visual question answering, or explanation generation [21, 42, 2, 47]. Although these methods demonstrate that text can be helpful, they mainly treat text as an auxiliary interface for improving classification, rather than explicitly investigating how real and generated images differ in the joint image–text space itself. This raises an important question: beyond visual artifacts, do generated images exhibit distinctive cross-modal structural properties relative to real images?

In this work, we answer this question from the perspective of image–text alignment. Intuitively, unlike real images captured from unconstrained physical scenes, generated images are synthesized under explicit or implicit conditioning signals, including constraints such as text. As a result, generated images may preserve target semantics more directly while containing fewer incidental visual factors unrelated to the conditioning content [30, 15, 36]. Motivated by this intuition, we conduct an empirical study in the joint image–text representation spaces of multimodal foundation models and observe a consistent phenomenon: compared with real images of similar semantics, generated images tend to exhibit stronger and more concentrated alignment with text. This phenomenon is reflected in multiple aspects, including smaller image–text residuals, more concentrated residual spectra, lower matching entropy, and a stronger tendency to be preferred in paired image–text retrieval, as shown in Fig. 1. We term this systematic discrepancy *Cross-modal Alignment Shift* (CAS).

Based on this observation, we further ask whether CAS can serve as a useful discriminative cue for AIGI detection. To this end, we propose a CAS-based detection framework that amplifies the generated-image-favored alignment response in multimodal models. Specifically, we introduce a text-guided alignment-shift amplification strategy to enlarge the CAS response in the joint image–text space. We then compare the representations produced by the original and amplified models for the same input image, and extract the resulting alignment-shift increment as an explicit CAS-aware discriminative signal. Finally, this increment is coupled with the amplified visual representation to construct a shift-aware feature for real/fake classification. In this way, the proposed framework transforms cross-modal alignment shift from an observed phenomenon into a discriminative representation for AIGI detection. Our contributions are summarized as follows:

- We reveal a systematic cross-modal alignment shift between real and generated images in vision-language representation spaces, quantified through pairwise retrieval preference, image–text residual structure, spectral concentration, and matching-distribution entropy.
- We further formulate CAS as a new cross-modal cue for AI-generated image detection, capturing the more regular, generated-favored alignment response induced by generative conditioning.
- We propose a text-guided alignment-shift amplification framework that enhances the generated-image-favored alignment response and extracts the resulting feature increment as an explicit cross-modal cue, coupled with amplified visual representations.

- Extensive experiments validate the effectiveness of the proposed method. In particular, when facing recent commercial generators, it outperforms prior state-of-the-art detectors by 5.48% and 6.13% in *ACC* and *AP*, respectively.

2 Related Work

2.1 AI-generated Image Detection

Existing AIGI detectors mainly exploit visual artifacts in the spatial, frequency, noise, or gradient domains [46, 11, 29, 40, 41]. Although effective on specific generators, these artifact-based cues often generalize poorly to unseen or high-quality generators. Recent methods therefore leverage pretrained vision–language models, especially CLIP, as universal feature extractors [31, 54, 51, 55]. However, these methods still mainly use the visual branch and leave the joint image–text representation structure underexplored.

2.2 Multimodal AI-Generated Image Detection

Beyond the image modality, existing studies have already explored using text to assist AI-generated image detection. DE-Fake [37] investigates how prompts used to generate forged images affect detection performance. CLIPping [21] shows, through a prompt-tuning adaptation strategy built on CLIP, that retaining the text branch helps improve performance. Antifake [3] formulates deepfake detection as a visual question answering task via prompt tuning. Zhou et al. [58] propose the Holmes Pipeline, which uses MLLMs to generate explanations for detection results. Although these methods use text to improve detection performance, they mainly treat text as a prompt interface and rarely explore the differences between real and generated images in the joint image–text representation space.

2.3 Alignment bias in AI-generated content

Some studies have explored the phenomenon that models tend to prefer AI-generated content. Dai et al. [8] found that information retrieval models tend to rank LLM-generated documents higher. Xu et al. [50] observed that image–text retrieval models tend to rank AI-generated images ahead of real images, and referred to this as invisible relevance bias. In addition, Huang et al. [17] found that LLM-generated content exhibits a smoother high-frequency distribution and steeper low-frequency components. While prior studies mainly view this preference as an undesirable retrieval bias to be diagnosed or mitigated, we repurpose it for AI-generated image detection by amplifying the generated-favored alignment response and extracting it as a discriminative cross-modal cue.

3 Cross-modal Alignment Shift: Phenomenon and Analysis

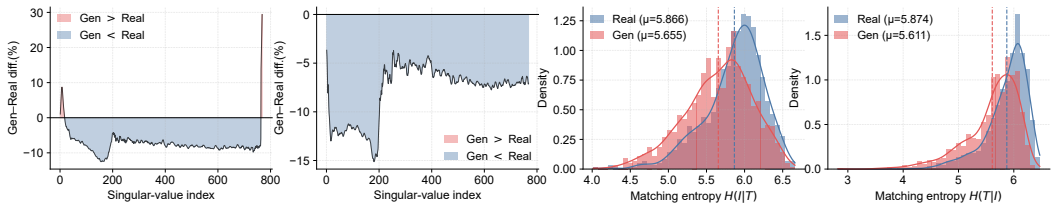
3.1 Paired Image–Text Construction

A central goal of this work is to reveal the difference between real and generated images in the joint image–text space. To minimize the effect of semantic mismatch, we construct paired samples whose real and generated images are closely matched in visual semantics. Specifically, we use Flickr30k [34] as the source of real images. Each image is associated with five captions, which we merge into a single description using ChatGPT and feed to a generative model to produce a synthetic counterpart.

To further reduce the semantic gap, we generate ten candidates for each real image using different random seeds, compute the cosine similarity between the real image and each candidate, and keep the most similar one as the final generated counterpart. Similarities are computed with the public CLIP ViT-L/14 image encoder on ℓ_2-normalized image embeddings; the selector uses no text and is not one of the analysis backbones. Unless otherwise stated, the analyses in this section use FLAVA [38] on 10,000 pairs (see Appendix D).

3.2 Uncovering Alignment Shift in Joint Vision-Language Space

To verify whether differences exist between real and generated images in the joint image–text space, we quantify the alignment preference of text toward real versus generated images in an image–text



(a) Image embedding spectrum (b) Image--text residual spectrum (c) T → I entropy (*d*=-0.50) (d) I → T entropy (*d*=-0.60)

Figure 2: **Empirical analysis of cross-modal alignment shift.** Generated images show more concentrated singular spectra, more compact image–text residuals, and lower matching entropy than real images. In (a,b), the y-axis reports the relative Gen–Real difference of normalized singular values. As further shown in Appendix A, these observations hold across multiple multimodal models, indicating that CAS is a general rather than model-specific phenomenon.

matching setting. For each text query, we match it against a mixed candidate pool composed of both the real image and the generated image, and examine which image is ranked first by the multimodal model. For the same text, we separately compute the average frequency with which its corresponding real image and generated image are ranked at the top. Intuitively, this reflects which type of image the text tends to align with more strongly when both real and generated images are present simultaneously. We use NDCG@1 over the mixed candidate pool to quantify this top-rank tendency. Specifically, we compute two NDCG@1 values in the same pool: M_{real}, where the paired real image is treated as the target, and M_{gen}, where the paired generated image is treated as the target. Since NDCG@1 is rank-based rather than similarity-scale-based, it enables comparison across multimodal models with different feature spaces. The normalized alignment discrepancy is then defined as

$$\text{NAD} = \frac{2\,(M_{\text{real}} - M_{\text{gen}})}{M_{\text{real}} + M_{\text{gen}}} \times 100\%. \tag{1}$$

Thus, NAD is the normalized difference between the real-target and generated-target NDCG@1 scores. $\text{NAD} > 0$ indicates that the model favors real images, whereas $\text{NAD} < 0$ indicates that the model favors generated images. Concretely, for N triplets each text query ranks one shared pool of all N real and N generated images; with a single target and a rank-1 cutoff, NDCG@1 is exactly the top-1 hit rate of that target, so NAD directly compares how often the paired real and the paired generated image are ranked first.

As shown in Fig. 1, cross-modal alignment shift is prevalent across a wide range of multimodal models. Even when real and generated images are highly similar in visual semantics, text still tends to align more strongly with generated images.

3.3 Understanding Cross-modal Alignment Shift

We further analyze why generated images exhibit stronger image–text alignment. Our hypothesis is that, since generated images are synthesized under conditioning signals, they tend to contain more concentrated target semantics and fewer incidental factors, thereby forming more regular correspondences with text.

First, we compare the singular-value spectra of real and generated image embeddings. Since larger singular values correspond to dominant representation components, a more concentrated spectrum indicates that the representation energy is carried by fewer principal directions. As shown in Fig. 2(a), generated images allocate more energy to the leading components than real images, with higher top-10 spectral energy (29.1% vs. 25.0%), lower effective dimension (115.4 vs. 128.5), and faster spectral decay. This suggests that generated images form more compact semantic structures in the representation space, whereas real images contain more dispersed factors of variation.

Second, we analyze the deviation between image and text representations. For each matched image–text pair, we define the image–text residual as

$$r_i = v(I_i) - t(T_i), \tag{2}$$

where $v(\cdot)$ and $t(\cdot)$ denote the visual and text encoders, respectively. The residual measures how much an image representation departs from its paired text representation. As shown in Fig. 2(b), residuals

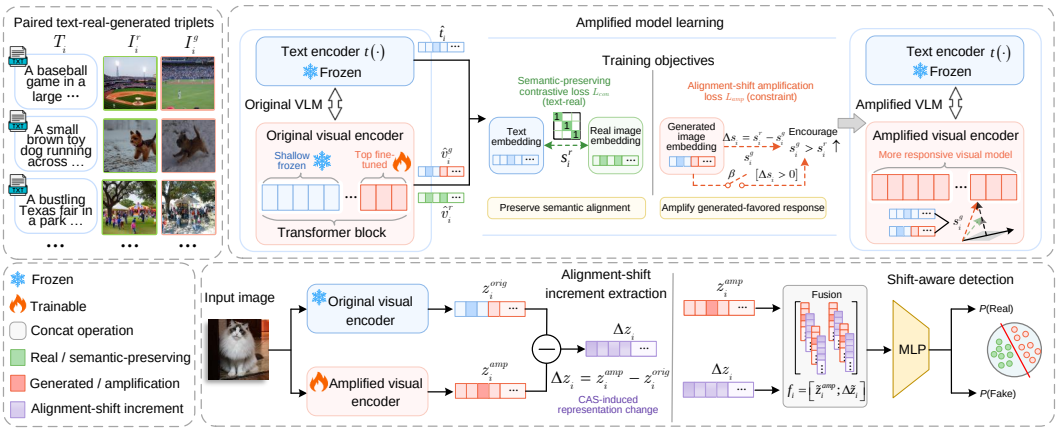


Figure 3: **Overview of the proposed CASNet.** CASNet learns an amplified vision–language encoder with a text-guided alignment-shift objective, estimates the amplification-induced increment by contrasting original and amplified visual representations, and constructs a shift-aware joint representation that couples the amplified visual response with the CAS increment for real/fake prediction.

of generated images are not only smaller in norm (1.093 vs. 1.148), but also more concentrated: their top singular component captures more residual energy (7.30% vs. 5.81%), and their effective dimension is lower (136.3 vs. 143.4). These results indicate that generated images deviate from text in a more regular and lower-dimensional manner than real images.

Finally, we examine matching certainty through the entropy of image–text matching distributions. Given a text query and a set of candidate images, lower entropy indicates a more peaked similarity distribution and thus less ambiguous matching. As shown in Fig. 2(c,d), generated images consistently yield lower entropy in both directions, including text-to-image matching (5.655 vs. 5.866, Cohen's $d = -0.50$) and image-to-text matching (5.611 vs. 5.874, Cohen's $d = -0.60$). This indicates that generated images establish more certain correspondences with text. The spectral and residual statistics use all 10,000 pairs, and the matching entropy uses a fixed 1,000-pair subset with temperature $\tau = 0.07$ (Appendix A).

4 Method

Building on the analysis in Section 3, our goal is to transform CAS from an observed cross-modal phenomenon into a usable detection signal. Since generated images show a stronger and more regular text-alignment response, we deliberately amplify this generated-favored response while preserving the original image–text semantic structure. The induced representation change is then extracted as a CAS increment and fused with the amplified visual feature, allowing the detector to exploit both cross-modal alignment evidence and visual discriminative cues. The overall framework, which we call CASNet, is illustrated in Fig. 3.

4.1 Learning an Amplified Model with an Image–Text Alignment Constraint

Given a training set of triplets (T_i, I_i^r, I_i^g), where T_i is the text description paired with the real image I_i^r and I_i^g is the corresponding generated image, we initialize from a pretrained vision–language model with text encoder $t(\cdot)$ and visual encoder $v(\cdot)$. During amplification, the text encoder is frozen and the visual encoder is fine-tuned to obtain $v_{\text{amp}}(\cdot)$. The goal is to preserve the basic image–text alignment ability while strengthening the cross-modal response that favors generated images.

Let $\hat{t}_i$, $\hat{v}_i^r$, and $\hat{v}_i^g$ denote the ℓ_2-normalized embeddings of $t(T_i)$, $v_{\text{amp}}(I_i^r)$, and $v_{\text{amp}}(I_i^g)$, respectively. To retain the original semantic alignment capability, we use a bidirectional contrastive loss on real image–text pairs. Let the similarity matrix be

$$S_{ij} = \tau \cdot \hat{t}_i^\top \hat{v}_j^r, \tag{3}$$

where τ is a scaling factor. The contrastive loss is defined as

$$\mathcal{L}_{con} = \frac{1}{2}\left(\mathcal{L}_{t\to i} + \mathcal{L}_{i\to t}\right), \tag{4}$$

with

$$\mathcal{L}_{t\to i} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{CE}\left(S_{i,:}, i\right), \quad \mathcal{L}_{i\to t} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{CE}\left(S_{:,i}, i\right). \tag{5}$$

This loss is computed only on real image–text pairs, and its role is to preserve the original joint representation capability of the pretrained model.

On top of this semantic-preserving objective, we introduce an alignment-shift amplification constraint. For each triplet, we compute the text–real and text–generated similarities:

$$s_i^r = \hat{t}_i^\top \hat{v}_i^r, \quad s_i^g = \hat{t}_i^\top \hat{v}_i^g, \tag{6}$$

and define their relative alignment gap as

$$\Delta s_i = s_i^r - s_i^g. \tag{7}$$

If $\Delta s_i > m$, the real image in this triplet exhibits a stronger image–text matching response than the generated image, where m is a response threshold and is set to 0 by default. This gap is used as a differentiable proxy for the rank-based NAD in Sec. 3.2. A smaller Δs_i indicates a stronger generated-favored alignment response. We impose the constraint only on samples that satisfy this condition, and further apply Bernoulli sampling with probability β to the eligible samples. The resulting target set is written as

$$B = \{i \mid \Delta s_i > m, b_i \sim \mathrm{Bernoulli}(\beta), b_i = 1\}. \tag{8}$$

The corresponding alignment-shift amplification loss is defined as

$$\mathcal{L}_{amp} = \begin{cases} \frac{1}{|B|}\sum\limits_{i\in B} \Delta s_i, & |B| > 0, \\ 0, & |B| = 0. \end{cases} \tag{9}$$

Minimizing this loss penalizes cases where real images align more strongly with text than generated images, thereby pushing the model to enhance the generated-image-favored alignment response. The overall objective is

$$\mathcal{L} = \mathcal{L}_{con} + \lambda_{amp}\mathcal{L}_{amp}, \tag{10}$$

where λ_{amp} balances semantic alignment preservation and alignment-shift amplification. Since $\mathcal{L}_{amp}$ averages over the sampled set B, β does not scale the conditional penalty magnitude: it subsamples eligible examples, which changes the update frequency (the loss is zero when B is empty) and the gradient variance, whereas λ_{amp} directly weights the loss and m determines eligibility. In practice, we freeze the text encoder and shallow visual layers, and only fine-tune the top visual Transformer blocks to avoid representation collapse.

4.2 Differential Extraction of Alignment-Shift Increments

After obtaining the amplified model, we extract a differential cue by comparing the original and amplified representations of the same image. Given an input image I_i, we denote the original and amplified visual features as

$$z_i^{\mathrm{orig}} = v_{\mathrm{orig}}(I_i), \quad z_i^{\mathrm{amp}} = v_{\mathrm{amp}}(I_i). \tag{11}$$

The alignment-shift increment is defined as

$$\Delta z_i = z_i^{\mathrm{amp}} - z_i^{\mathrm{orig}}. \tag{12}$$

This increment characterizes how the representation of the same image changes after alignment-shift amplification. Since the amplified encoder is optimized to emphasize the generated-image-favored cross-modal response, Δz_i highlights the CAS-induced representation change while suppressing factors shared by the original and amplified encoders. Therefore, it serves as an explicit differential cue for modeling cross-modal alignment discrepancy.

Table 1: Detection performance in terms of Acc.(%) and A.P.(%) on diffusion generators. **Bold** indicates the best result in each column. Underline indicates the second best.

Method	SDV1.4		VQDM		SDV1.5		GLIDE		ADM		FLUX		PixArt-XL		SD35-L		Mean	
	ACC	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*
Fusing [18]	99.90	99.80	64.20	84.60	99.91	97.98	57.50	94.20	51.30	69.30	59.04	91.86	73.31	99.55	58.29	97.53	70.43	91.85
UnivFD [31]	86.60	96.20	60.70	78.40	87.80	96.26	78.20	93.73	58.85	72.25	77.00	94.43	80.00	97.82	68.50	91.15	74.71	90.03
NPR [41]	**100.00**	**100.00**	68.30	86.30	99.90	99.97	89.70	95.80	73.00	94.70	82.00	**99.80**	91.00	**100.00**	71.00	99.20	84.36	96.97
DRCT [4]	95.01	99.65	**90.03**	96.54	94.41	99.49	89.18	94.64	**79.42**	88.96	81.20	90.18	86.20	91.25	87.60	93.51	87.88	94.28
CLIPping [21]	96.07	93.97	79.35	77.17	95.48	93.10	85.00	87.44	70.14	68.00	70.89	70.51	84.20	84.00	94.39	94.30	84.44	83.56
CO-SPY [6]	88.88	95.34	78.53	88.65	94.25	98.63	80.78	91.63	68.67	82.42	62.10	62.75	71.40	78.45	73.70	79.44	77.29	84.66
Effort [51]	**100.00**	**100.00**	89.50	**99.71**	**100.00**	**100.00**	85.00	97.38	72.00	91.00	70.45	96.10	86.95	99.50	93.45	**99.77**	87.17	97.93
VIB-Net [54]	99.55	**100.00**	89.35	97.00	99.20	99.97	74.25	97.13	73.85	95.49	**82.55**	96.87	89.40	98.73	93.10	99.17	87.66	98.05
Ours	99.74	**100.00**	86.94	99.27	99.60	99.99	**96.02**	**99.74**	78.94	**98.18**	81.85	96.54	**95.54**	99.79	**95.18**	99.55	**91.73**	**99.13**

Table 2: Detection performance in terms of Acc.(%) and A.P.(%) on GAN and autoregressive.

Method	ProGAN		StyleGAN		CycleGAN		StarGAN		BigGAN		GauGAN		DFGAN		Infinity		LlamaGen		Show-o		Mean	
	ACC	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*
Fusing [18]	51.20	67.63	50.20	67.90	52.40	87.79	58.20	91.20	53.50	69.37	49.32	43.08	90.90	85.26	95.84	99.96	73.13	99.63	82.26	99.53	65.70	81.14
UnivFD [31]	61.75	58.37	66.50	59.20	65.55	67.22	67.20	59.93	65.65	80.21	60.20	70.90	61.50	84.76	88.00	98.71	79.50	96.17	83.00	98.40	69.89	77.39
NPR [41]	60.30	83.30	58.00	82.70	67.20	94.90	73.20	97.30	59.20	72.00	52.00	66.00	59.00	96.70	95.00	**100.00**	95.00	**100.00**	**98.00**	**100.00**	71.69	89.29
DRCT [4]	76.83	91.03	71.58	79.44	**94.51**	98.68	63.13	96.29	82.68	93.51	78.39	86.62	93.10	98.88	94.80	99.07	92.70	96.26	82.20	89.82	82.99	92.96
CLIPping [21]	88.78	88.54	80.69	77.23	88.48	88.44	90.82	89.82	89.57	85.33	85.94	81.56	98.00	98.80	99.00	98.70	93.50	93.10	87.90	87.70	90.27	88.92
CO-SPY [6]	77.84	82.26	68.88	69.95	69.65	68.28	67.48	71.40	76.75	78.41	77.57	78.74	92.30	98.39	81.90	93.92	88.60	95.38	84.60	94.45	78.56	83.12
Effort [51]	92.90	98.58	75.00	95.61	88.50	92.60	95.50	98.43	**93.50**	**97.56**	80.00	89.93	98.00	**100.00**	99.00	99.98	95.75	99.80	91.30	99.49	90.95	**97.20**
VIB-Net [54]	89.70	96.59	74.10	84.31	88.60	98.44	80.70	97.60	91.20	97.17	**87.15**	**96.94**	**99.00**	99.85	**99.10**	99.97	**96.55**	99.63	92.35	99.20	89.85	96.97
Ours	**94.77**	**98.90**	**92.41**	**98.22**	79.17	**98.80**	**99.22**	**99.96**	82.77	90.36	77.06	86.11	98.05	99.98	99.02	99.95	93.45	99.69	93.75	99.00	**90.97**	97.10

Table 3: Detection performance in terms of Acc.(%) and A.P.(%) on recent commercial generators.

Method	Imagen-3		DALL·E 2		DALL·E 3		Midjourney		Hourglass		IdeogramV2		FLUX-schnell		FLUX-dev		Wukong		Mean	
	ACC	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*	*ACC*	*AP*
Fusing [18]	85.71	99.04	53.90	94.06	98.80	99.94	52.30	81.20	55.50	79.86	83.30	99.23	60.10	94.23	71.40	97.44	99.90	99.90	73.43	93.88
UnivFD [31]	67.50	75.92	47.00	53.38	72.00	80.31	75.90	92.08	54.00	57.40	63.50	61.73	70.00	69.87	84.00	88.63	78.20	90.98	68.01	74.48
NPR [41]	87.00	99.50	54.00	77.00	**100.00**	**100.00**	82.30	95.50	76.00	**85.70**	67.00	81.60	**87.00**	**99.80**	86.00	91.50	**100.00**	**100.00**	82.14	92.29
DRCT [4]	79.50	85.16	78.30	83.24	82.30	95.70	**88.50**	97.03	54.20	60.07	65.10	65.98	67.40	65.94	70.50	76.27	94.68	99.37	75.61	80.97
CLIPping [21]	95.50	94.82	88.70	89.21	97.00	97.22	77.66	77.34	52.80	51.46	85.40	86.36	59.30	58.22	68.56	67.33	88.86	86.52	79.31	78.72
CO-SPY [6]	86.80	95.40	81.20	89.75	85.60	95.39	78.59	97.81	55.50	65.92	69.90	70.11	69.20	69.16	56.89	52.97	88.93	99.93	74.73	81.83
Effort [51]	94.95	99.06	**95.00**	**99.38**	**100.00**	**100.00**	78.00	**98.68**	46.00	51.64	**91.50**	**99.82**	69.50	91.34	59.50	84.21	99.50	**100.00**	81.55	91.57
VIB-Net [54]	88.30	95.39	84.00	91.77	95.50	99.65	88.05	97.81	54.00	56.47	84.50	93.85	72.50	82.14	78.00	87.31	98.25	99.93	82.57	89.37
Ours	**97.27**	**99.65**	89.19	98.78	97.07	98.78	74.94	97.03	**77.33**	71.55	86.02	98.36	82.59	96.86	**89.34**	**98.54**	98.65	99.93	**88.05**	**95.50**

4.3 Shift-Aware Joint Representation for Detection

Although Δz_i captures the amplification-induced cross-modal change, using it alone may discard useful visual evidence, such as texture artifacts, local inconsistencies, and semantic irregularities. We therefore jointly model two complementary signals: the amplified visual feature, which retains discriminative visual responses, and the alignment-shift increment, which captures the CAS-specific differential response.

We denote the ℓ_2-normalized amplified feature and alignment-shift increment as $\tilde{z}_i^{\mathrm{amp}}$ and $\Delta\tilde{z}_i$, respectively. The final shift-aware representation is constructed as

$$f_i = [\tilde{z}_i^{\mathrm{amp}}; \Delta\tilde{z}_i]. \tag{13}$$

Here, $\tilde{z}_i^{\mathrm{amp}}$ provides the amplified visual response, while $\Delta\tilde{z}_i$ explicitly encodes the CAS-induced representation variation between the original and amplified models. The concatenated representation f_i is then fed into a lightweight MLP classifier for real/fake prediction. This design allows CASNet to preserve conventional visual discriminative cues while explicitly incorporating structural evidence from the joint image–text space.

5 Experiments

5.1 Experimental Settings

Training Dataset. We use 10K Flickr image–text pairs, consisting of 10K real images and their SDv1.5-generated counterparts, to train the amplified model. Following prior work [54, 51, 41], the detector is trained on the Stable Diffusion v1.4 subset of GenImage for fair comparison.

Test Datasets. We evaluate our method on four benchmarks: GenImage [60], ForenSynths [46], AIGI-Holmes [58], and Community Forensics [32]. These datasets cover diverse generation paradigms, including diffusion models, GANs, autoregressive models, and commercial text-to-image systems. Specifically, the diffusion models include SDV1.4 [35], VQDM [13], SDV1.5 [35], GLIDE [30], ADM [9], FLUX [22], PixArt-XL [5], and SD35-L. The GAN and autoregressive models include BigGAN [1], CycleGAN [59], GauGAN [33], ProGAN [19], StarGAN [7], StyleGAN [20], DFGAN [43], Infinity [14], LlamaGen [39], and Show-o [48], and the commercial models include Imagen-3, DALL-E 2, DALL-E 3, Midjourney, Hourglass, IdeogramV2, FLUX-schnell, FLUX-dev, and Wukong.

Implementation Details. We implement all experiments in PyTorch on NVIDIA H100 PCIe GPUs. DFNCLIP ViT-L/14 [10] is used as the default backbone and initialized with publicly available pretrained weights. During amplification training, the text encoder and the first 20 visual Transformer layers are frozen, while only the upper visual layers are fine-tuned. We use AdamW with a learning rate of 1×10^{-6}, weight decay of 0.01, batch size of 16, 50 epochs, and a warm-up ratio of 0.1. The hyperparameters are set to $\beta = 0.5$, $m = 0$, and $\lambda_{\text{amp}} = 1.0$. The downstream detector is a single-hidden-layer MLP trained with Adam using a learning rate of 1×10^{-3} and a batch size of 256. We also profile CASNet in terms of parameters, GFLOPs, inference latency, per-epoch training time, and GPU memory, detailed results are provided in Appendix B.

Baseline Methods. We compare our method with a range of state-of-the-art detection approaches, including Fusing (ICIP 2022) [18], UnivFD (CVPR 2023) [31], CLIPping (ICMR 2024) [21], NPR (CVPR 2024) [41], DRCT (ICML 2024) [4], CO-SPY (CVPR 2025) [6], VIB-Net (CVPR 2025) [54], and Effort (ICML 2025) [51]. To ensure fair comparison, all methods are conducted under the same experimental settings, including both training and testing.

5.2 Performance Comparisons

In this section, we compare our method with a series of state-of-the-art generated-image detection methods across multiple benchmark datasets.

Tables 1 and 2 report the results on diffusion, GAN, and autoregressive generators. Multimodal-model-based detectors such as UnivFD and VIB-Net mainly rely on visual artifacts and therefore degrade on advanced generators with weak artifacts, while CLIPping uses text primarily as an adaptation interface rather than modeling real–fake differences in the joint image–text space, leading to unstable performance across benchmarks.

In contrast, our method achieves the best average performance on the diffusion and commercial suites and is competitive with the strongest baselines on GAN and autoregressive generators, although it is not the best on every individual generator. In particular, it improves *ACC* on diffusion benchmarks by 4.1% over VIB-Net. On recent commercial models (Table 3), it improves *ACC* and *AP* over VIB-Net by 5.48% and 6.13%, respectively. These results show that cross-modal alignment shift remains effective even when visual artifacts are weak, underscoring the value of modeling structural differences in the joint image–text space.

5.3 Ablation Study

Effect of Individual Components. Table 4 summarizes the ablation results. Compared with the baseline, the amplified feature improves the average accuracy from 85.38% to 87.73% and *AP* from 95.28 to 97.02, showing that alignment-shift amplification strengthens the discriminative response, especially on diffusion models. Using only the differential increment degrades performance on GANs, likely because differencing suppresses not only CAS-related information but also useful visual artifact cues. In contrast, combining the amplified feature with the shift increment yields the best overall performance, raising the average accuracy to 90.48% and *AP* to 97.69%, which indicates that CAS and visual cues are complementary.

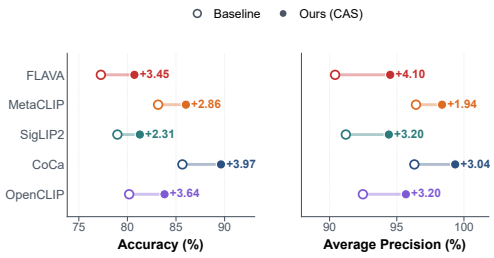


Figure 4: **Effect of transferring CAS to different multimodal backbones.**

Table 4: Ablation study on representation design.

Design	Representation	Diff.-set		GAN-set		Mean	
		Acc.	*A.P.*	*Acc.*	*A.P.*	*Acc.*	*A.P.*
Baseline	z^{orig}	85.04	95.78	85.78	94.70	85.38	95.28
Amplified response	z^{amp}	88.64	98.92	86.68	94.85	87.73	97.02
Differential increment	$\Delta z = z^{\text{amp}} - z^{\text{orig}}$	87.41	98.41	80.49	90.53	84.18	94.73
CASNet	$\text{Concat}(z^{\text{amp}}, \Delta z)$	**91.73**	**99.13**	**89.06**	**96.05**	**90.48**	**97.69**

Table 5: Effect of generator choice for constructing the amplified alignment response.

Model	Diff.-set		GAN-set		Mean	
	Acc.	*A.P.*	*Acc.*	*A.P.*	*Acc.*	*A.P.*
SDXL	91.23	99.11	89.13	96.13	90.25	97.72
VQDM	91.02	99.05	**89.50**	**96.28**	90.31	**97.76**
FLUX	91.12	99.06	89.11	**96.28**	90.18	**97.76**
SDv1.5	**91.73**	**99.13**	89.06	96.05	**90.48**	97.69

Table 6: Effect of the number of samples used to fit the alignment-amplification model.

#Samples	Diff.-set		GAN-set		Mean	
	Acc.	*A.P.*	*Acc.*	*A.P.*	*Acc.*	*A.P.*
30k	90.09	99.02	86.55	94.38	88.44	96.85
10k	**91.73**	**99.13**	89.06	96.05	90.48	97.69
5k	91.37	**99.13**	89.22	**96.30**	90.37	**97.81**
1k	91.62	**99.13**	**89.74**	96.10	**90.74**	97.71

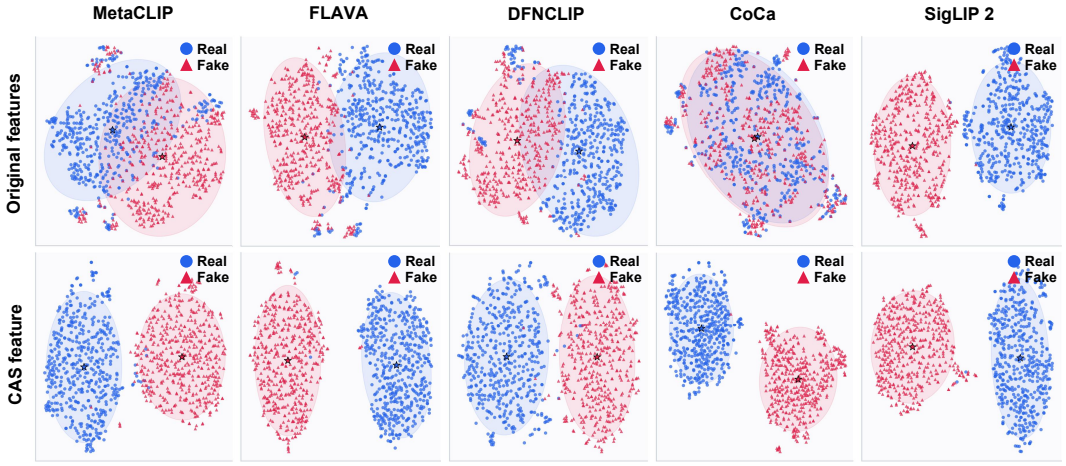


Figure 5: **Feature visualization on SDv1.5.** Across five multimodal backbones, CAS features produce more compact and better separated real/fake clusters than the original features.

Choice of Generative Model for Training the Amplified Model. Table 5 shows that using SDXL, VQDM, FLUX, or SDv1.5 to train the amplified model yields similar performance. This suggests that CAS is not tied to a specific generator and can be amplified from different generated-image sources.

Amount of Training Data for Alignment-Shift Amplification. Table 6 shows that our method remains effective even with only 1k–5k generated images. This indicates that the alignment-shift amplification does not require large-scale training data.

Different Backbone Models. To evaluate the adaptability and generality of our method, we further apply it to other multimodal models, including FLAVA, MetaCLIP, SigLIP2, CoCa, and OpenCLIP. The results in Fig. 4 show that our method can be seamlessly applied to a wide range of multimodal models and consistently improve their detection performance.

Feature Visualization Analysis. We use t-SNE [45] to visualize feature distributions on SDv1.5 and assess the effect of CAS. As shown in Fig. 5, the original features are often mixed, whereas the learned CAS representation yields more compact and clearly separated real/fake clusters across all backbones, indicating that the amplification transfers across multimodal models.

5.4 Robustness Evaluation

We further evaluate robustness under common image post-processing operations, including downscaling, Gaussian noise, color jitter, JPEG compression, and Gaussian blur. As shown in Appendix C, CASNet achieves the best accuracy across all 15 perturbation settings and the highest average accuracy of 76.44%, outperforming the strongest baseline by 7.54 percentage points. These results indicate that CASNet provides a stable detection cue under practical image degradations.

6 Conclusion

In this paper, we studied AI-generated image detection from the perspective of image–text alignment. We revealed a consistent phenomenon termed Cross-modal Alignment Shift (CAS), where generated images tend to align with text more strongly and more regularly than semantically matched real images. This shift is supported by retrieval preference, image–text residuals, spectral concentration, and matching entropy, indicating that generated images exhibit a more concentrated alignment pattern in the joint vision–language space.

Based on CAS, we proposed a detection framework (CASNet) that amplifies the generated-image-favored alignment response and extracts the resulting feature increment as an explicit cross-modal cue. By coupling this increment with amplified visual representations, the detector exploits both alignment-level structural evidence and visual discriminative cues. Experiments across diffusion, GAN, autoregressive, and commercial generators demonstrate the effectiveness and generality of CAS for AI-generated image detection.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers and the area chair for their constructive comments, which helped us improve this paper. Funding in direct support of this work: the National Natural Science Foundation of China under Grant 62376064, Grant 62536002, Grant 62561160098, and Grant 62576060; the National Natural Science Foundation of China Joint Key Project under Grant U24B20182 and Grant U24B20173; the Natural Science Foundation of Chongqing under Grant CSTB2025YITP-QCRCX0074; and the Chongqing Graduate Research and Innovation Project under Grant CYB25261. Competing interests: the authors declare no financial relationships outside the affiliations listed above that could be perceived as influencing this work.

References

[1] Andrew Brock. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[2] George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10759–10769, 2024.

[3] 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.

[4] 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.

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

[6] Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13455–13465, 2025.

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[8] Sunhao Dai, Yuqi Zhou, Liang Pang, Weihao Liu, Xiaolin Hu, Yong Liu, Xiao Zhang, Gang Wang, and Jun Xu. Neural retrievers are biased towards llm-generated content. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 526–537, 2024.

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[10] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

[11] 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*, pages 3247–3258. PMLR, 2020.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[13] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022.

[14] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15733–15744, 2025.

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

[17] Wei Huang, Keping Bi, Yinqiong Cai, Wei Chen, Jiafeng Guo, and Xueqi Cheng. How do llm-generated texts impact term-based retrieval models? In *Proceedings of the Nineteenth ACM International Conference on Web Search and Data Mining*, pages 238–248, 2026.

[18] 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)*, pages 3465–3469. IEEE, 2022.

[19] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[21] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1006–1015, 2024.

[22] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

[23] Shuaibo Li, Shibiao Xu, Wei Ma, and Qiu Zong. Image manipulation localization using attentional cross-domain cnn features. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5614–5628, 2021.

[24] Shuaibo Li, Wei Ma, Jianwei Guo, Shibiao Xu, Benchong Li, and Xiaopeng Zhang. Unionformer: Unified-learning transformer with multi-view representation for image manipulation detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12523–12533, 2024.

[25] Shuaibo Li, Zhaohu Xing, Hongqiu Wang, Pengfei Hao, Xingyu Li, Zekai Liu, and Lei Zhu. Toward medical deepfake detection: A comprehensive dataset and novel method. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 626–637. Springer, 2025.

[26] Shuaibo Li, Pengfei Hao, Hongtao Wu, Jianfeng Dong, Ping Li, Xiaohong Liu, and Lei Zhu. Hermes: An evidence-driven agentic framework for trustworthy and explainable ai-generated video detection. In *Forty-third International Conference on Machine Learning*, 2026.

[27] Shuaibo Li, Yijun Yang, Zhaohu Xing, Hongqiu Wang, Pengfei Hao, Xingyu Li, Zekai Liu, Qing Zhang, and Lei Zhu. Synerdetect: Hierarchical synergistic learning for generalizable ai-generated image detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6405–6414, 2026. doi: 10.1609/aaai.v40i8.37568.

[28] Shuaibo Li, Laixin Zhang, Wei Ma, Jianwei Guo, Shibiao Xu, Zhijie Qiu, and Hongbin Zha. Realnet: Efficient and unsupervised detection of ai-generated images via real-only representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38889–38898, 2026.

[29] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *European Conference on Computer Vision*, pages 95–110. Springer, 2022.

[30] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[31] 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*, pages 24480–24489, 2023.

[32] Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8245–8257, 2025.

[33] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019.

[34] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

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

[37] 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*, pages 3418–3432, 2023.

[38] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650, 2022.

[39] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

[40] 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*, pages 12105–12114, 2023.

[41] 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*, pages 28130–28139, 2024.

[42] 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, pages 7184–7192, 2025.

[43] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16515–16525, 2022.

[44] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[45] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[46] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020.

[47] 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*.

[48] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

[49] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.

[50] Shicheng Xu, Danyang Hou, Liang Pang, Jingcheng Deng, Jun Xu, Huawei Shen, and Xueqi Cheng. Invisible relevance bias: Text-image retrieval models prefer ai-generated images. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 208–217, 2024.

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

[52] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[53] Lin Yuan, Xiaowan Li, Yan Zhang, Jiawei Zhang, Hongbo Li, and Xinbo Gao. Mlep: Multi-granularity local entropy patterns for generalized ai-generated image detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[54] Haifeng Zhang, Qinghui He, Xiuli Bi, Weisheng Li, Bo Liu, and Bin Xiao. Towards universal ai-generated image detection by variational information bottleneck network. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23828–23837, 2025.

[55] Haifeng Zhang, Qinghui He, Xiuli Bi, Bo Liu, Chi-Man Pun, and Bin Xiao. Combating pattern and content bias: Adversarial feature learning for generalized ai-generated image detection. *IEEE Transactions on Image Processing*, 2026. doi: 10.1109/TIP.2026.3742643.

[56] Laixin Zhang, Shuaibo Li, Wei Ma, and Hongbin Zha. Truemoe: Dual-routing mixture of discriminative experts for synthetic image detection. *arXiv preprint arXiv:2509.15741*, 2025.

[57] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021.

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

[59] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[60] 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 Additional Analysis of Cross-modal Alignment Shift Across Backbones

Section 3 analyzes Cross-modal Alignment Shift (CAS) using one representative vision–language backbone. In this appendix, we provide additional evidence that the observed phenomenon is not tied to a particular multimodal representation space. We repeat the same analysis protocol on three additional dual-tower vision–language encoders: SigLIP2, DFNCLIP, and CoCa. These models differ in their pretraining data, architecture design, and contrastive or generative vision–language objectives, making them suitable for testing whether CAS is a backbone-specific artifact or a more general structural property.

Data and feature extraction. We use the same paired image–text construction as in Section 3.1. Each sample is a triplet (T_i, I_i^r, I_i^g), where T_i is the text description, I_i^r is the real image, and I_i^g is the corresponding generated image constructed to be semantically close to the real image. For each backbone, we extract an image embedding and a text embedding using its visual and textual encoders. All embeddings are ℓ_2-normalized before analysis:

$$t_i = \frac{f_t(T_i)}{\|f_t(T_i)\|_2}, \qquad v_i^r = \frac{f_v(I_i^r)}{\|f_v(I_i^r)\|_2}, \qquad v_i^g = \frac{f_v(I_i^g)}{\|f_v(I_i^g)\|_2}. \tag{14}$$

Here, $f_t(\cdot)$ and $f_v(\cdot)$ denote the text and image encoders of the corresponding backbone. Unless otherwise stated, the SVD-based statistics and residual-norm statistics are computed over all $N = 10{,}000$ paired samples. For matching-entropy analysis, we follow the same implementation as in the main analysis and estimate the entropy on a fixed random subset of 1,000 paired samples for efficient computation of the full pairwise similarity matrix.

A.1 Spectral Analysis of Image Embeddings

We first examine whether real and generated images occupy representation spaces with different degrees of concentration. Let

$$V^r = [v_1^r; v_2^r; \cdots; v_N^r] \in \mathbb{R}^{N \times d}, \qquad V^g = [v_1^g; v_2^g; \cdots; v_N^g] \in \mathbb{R}^{N \times d} \tag{15}$$

be the real and generated image embedding matrices. Before singular value decomposition, we center each matrix by subtracting its column-wise mean. For a centered embedding matrix X, let its singular values be $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$. We define the normalized spectral energy of the k-th component as

$$p_k = \frac{\sigma_k^2}{\sum_{j=1}^{d} \sigma_j^2}. \tag{16}$$

The top-K spectral energy is then

$$E_K = \sum_{k=1}^{K} p_k. \tag{17}$$

A larger E_K means that more representation energy is carried by the leading principal directions, indicating a more concentrated representation structure.

We also compute the effective rank:

$$r_{\text{eff}} = \exp\left(-\sum_{k=1}^{d} p_k \log p_k \right). \tag{18}$$

The effective rank can be interpreted as the number of principal directions that effectively contribute to the representation. A smaller effective rank indicates that the representation lies in a more compact, lower-dimensional subspace. In our implementation, we additionally record the stable rank and participation ratio:

$$r_{\text{stable}} = \frac{\sum_{k=1}^{d} \sigma_k^2}{\sigma_1^2}, \qquad r_{\text{PR}} = \frac{\left(\sum_{k=1}^{d} \sigma_k^2\right)^2}{\sum_{k=1}^{d} \sigma_k^4}. \tag{19}$$

These two quantities provide complementary diagnostics of spectral concentration. The compact summary in Table 7 primarily reports top-10 spectral energy and effective rank, since they directly capture leading-component dominance and overall dimensionality.

For visualization, Fig. 6 plots the relative singular-value difference between generated and real images. Specifically, we normalize the singular values of each group by its leading singular value, $\hat{\sigma}_k = \sigma_k / \sigma_1$, and plot

$$\Delta_{\text{SV}}(k) = 100 \times \frac{\hat{\sigma}_k^g - \hat{\sigma}_k^r}{\hat{\sigma}_k^r + \epsilon}. \tag{20}$$

Positive regions indicate that generated images have larger normalized singular values at the corresponding index, while negative regions indicate the opposite. Together with the top-K energy and effective rank, this plot describes how the spectral shape differs between real and generated image embeddings.

A.2 Image–Text Residual Analysis

We next analyze how image embeddings deviate from their paired text embeddings. For each real and generated image–text pair, we define the image–text residuals as

$$r_i^r = v_i^r - t_i, \qquad r_i^g = v_i^g - t_i. \tag{21}$$

The residual vector measures the part of the image representation that is not explained by the paired text representation. If generated images are more directly aligned with their conditioning semantics, their residuals should be smaller or more regular than those of real images.

We form residual matrices

$$R^r = [r_1^r; r_2^r; \cdots; r_N^r], \qquad R^g = [r_1^g; r_2^g; \cdots; r_N^g], \tag{22}$$

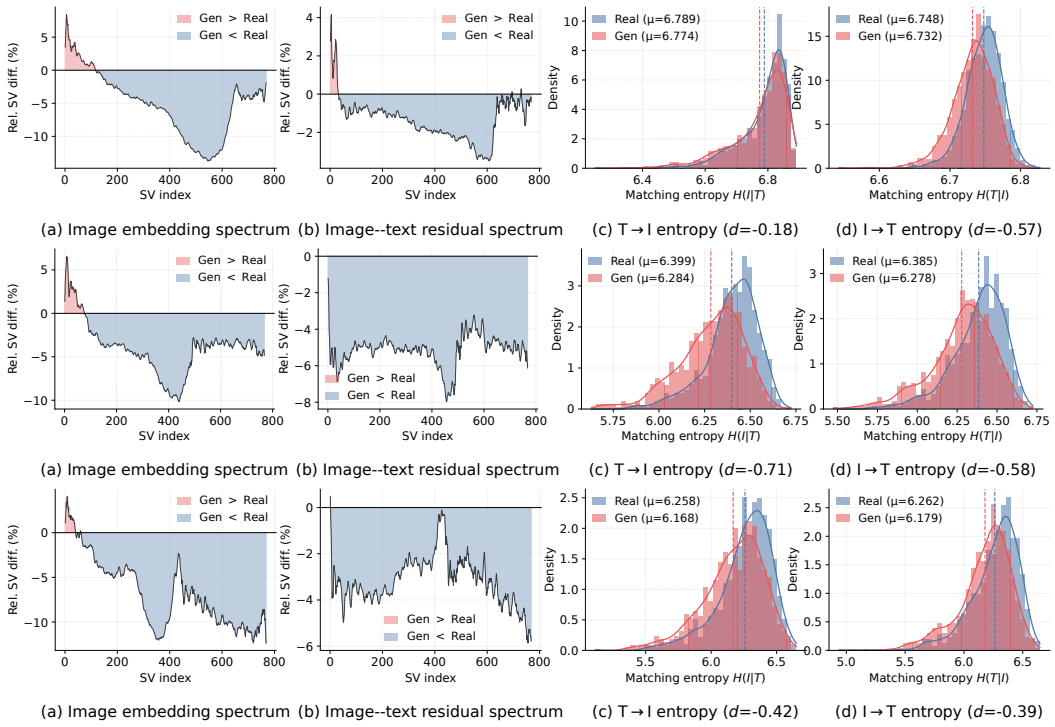


Figure 6: Additional empirical analysis of CAS across multimodal backbones. From top to bottom, the rows correspond to SigLIP2, DFNCLIP, and CoCa. Each row follows the same protocol as Fig. 2: image-embedding singular spectrum, image–text residual singular spectrum, text-to-image matching entropy, and image-to-text matching entropy. Across all three backbones, generated images exhibit more concentrated spectra and lower matching entropy than real images, confirming that CAS is not specific to the backbone used in the main analysis.

and apply the same SVD-based analysis as above. Higher top-K residual energy and lower residual effective rank indicate that the image–text deviations are more concentrated in fewer principal residual directions. We also compute the mean residual norm:

$$\bar{d}^r_{\text{res}} = \frac{1}{N}\sum_{i=1}^{N} \|r^r_i\|_2, \qquad \bar{d}^g_{\text{res}} = \frac{1}{N}\sum_{i=1}^{N} \|r^g_i\|_2. \tag{23}$$

A smaller residual norm means that the image embedding is closer to its paired text embedding in the normalized joint representation space. Therefore, if $\bar{d}^g_{\text{res}} < \bar{d}^r_{\text{res}}$, generated images have smaller average image–text deviations than real images.

A.3 Matching-Distribution Entropy

Finally, we evaluate the certainty of image–text matching distributions. For text-to-image matching, we take a text query t_i and compare it with a candidate image set from either real images or generated images. For $c \in \{r, g\}$, the temperature-scaled similarity is

$$a^c_{ij} = \frac{t_i^\top v^c_j}{\tau}, \tag{24}$$

where τ is the temperature parameter. The matching distribution over candidate images is obtained by a softmax:

$$p^{c,T\to I}_{ij} = \frac{\exp(a^c_{ij})}{\sum_{l=1}^{S}\exp(a^c_{il})}. \tag{25}$$

Table 7: Quantitative summary of additional CAS analysis across multimodal backbones. Each entry reports Real / Generated values. Generated images consistently show higher leading spectral energy, lower effective rank, smaller residual norm, and lower matching entropy.

Backbone	Img. Top-10 ↑	Img. Eff. Rank ↓	Res. Top-10 ↑	Res. Norm ↓	$H(I\|T)$ ↓	$H(T\|I)$ ↓
SigLIP2	28.24 / 30.44	183.9 / 164.5	21.31 / 22.42	1.327 / 1.319	6.789 / 6.774	6.748 / 6.732
DFNCLIP	26.56 / 28.20	183.7 / 168.0	17.83 / 18.45	1.190 / 1.180	6.399 / 6.284	6.385 / 6.278
CoCa	30.81 / 32.22	138.3 / 128.4	20.85 / 21.54	1.111 / 1.101	6.258 / 6.168	6.262 / 6.179

The text-to-image matching entropy is

$$H^c(I|T) = \frac{1}{S}\sum_{i=1}^{S}\left(-\sum_{j=1}^{S} p_{ij}^{c,T\to I}\log p_{ij}^{c,T\to I}\right). \tag{26}$$

A lower value of $H(I|T)$ means that the text query induces a sharper and less ambiguous distribution over candidate images.

Similarly, for image-to-text matching, we take an image query v_i^c and compare it with the candidate text set. The similarity, distribution, and entropy are defined as

$$b_{ij}^c = \frac{(v_i^c)^\top t_j}{\tau}, \tag{27}$$

$$p_{ij}^{c,I\to T} = \frac{\exp(b_{ij}^c)}{\sum_{l=1}^{S}\exp(b_{il}^c)}, \tag{28}$$

and

$$H^c(T|I) = \frac{1}{S}\sum_{i=1}^{S}\left(-\sum_{j=1}^{S} p_{ij}^{c,I\to T}\log p_{ij}^{c,I\to T}\right). \tag{29}$$

Lower $H(T|I)$ indicates that an image query yields a more certain distribution over candidate texts. In addition to entropy, our implementation records the ground-truth top-1 probability p_{ii} and the matching margin $a_{ii} - \max_{j\neq i} a_{ij}$ or $b_{ii} - \max_{j\neq i} b_{ij}$ as auxiliary diagnostics. Higher top-1 probability and larger margin indicate a more confident matched pair. Since entropy already summarizes the sharpness of the full matching distribution, we report entropy as the main matching-certainty metric.

For the entropy histograms in Fig. 6, we also report Cohen's d as an effect-size statistic:

$$d = \frac{\mu_g - \mu_r}{s_{\text{pooled}}}, \qquad s_{\text{pooled}} = \sqrt{\frac{(n_r-1)s_r^2 + (n_g-1)s_g^2}{n_r+n_g-2}}. \tag{30}$$

Here, μ_r, μ_g and s_r, s_g are the means and standard deviations of the real and generated entropy distributions. A negative value of d means that generated images have lower matching entropy than real images.

A.4 Results Across Additional Backbones

Fig. 6 and Table 7 summarize the additional analysis. Across SigLIP2 [44], DFNCLIP, and CoCa [52], generated images consistently exhibit higher top-10 image spectral energy and lower image effective rank than real images. This indicates that generated image embeddings allocate more energy to leading representation directions and form a more compact visual-semantic structure.

The residual analysis shows a similar trend. Generated image–text residuals have higher top-10 residual spectral energy and lower residual effective rank across the three backbones. Their mean residual norm is also smaller than that of real images, suggesting that generated images deviate less from their paired text embeddings and that the remaining deviations are organized in a more regular, lower-dimensional manner.

Table 8: Computational cost comparison. GFLOPs are reported under the MACs-equivalent convention used by the baseline resource table. For CASNet, training time is reported under the online-equivalent setting that recomputes frozen visual features every epoch.

Method	Parameters		GFLOPs	Inference Time/ms	Training Time/Epoch(s)	Memory Usage(MB)
	Total(M)	Trainable				
CNNSpot	25.56	25,557,032	4.14	4.74	9034.49	332.49
UnivFD	303.97	769	77.83	10.92	8748.65	1268.62
CLIPping	389.70	24,576	90.93	45.46	9889.62	1662.65
VIB-Net	306.33	2,362,113	77.83	11.37	7248.75	1278.10
CASNet	608.39	459,650	155.55	20.62	6687.09	3698.50

The matching-entropy results further support this interpretation. In both text-to-image and image-to-text directions, generated images yield lower entropy than real images across all three backbones. The magnitude of the entropy shift varies across models: for SigLIP2 the difference is relatively small, while DFNCLIP and CoCa show more pronounced entropy reductions. Nevertheless, the direction of the effect is consistent. These results support the conclusion that CAS reflects a general structural difference between real and generated images in joint image–text representation spaces, rather than an artifact of the single backbone used in the main analysis.

B Computational Cost Analysis

Purpose. Since CASNet detects generated images by comparing the responses of the original and amplified visual encoders, it introduces an additional frozen visual forward pass compared with single-backbone detectors. We therefore conduct a computational cost analysis to examine whether this extra CAS modeling cost leads to impractical training or inference overhead.

Evaluation protocol. We compare CASNet with representative AIGI detectors using six metrics: total parameters, trainable parameters, GFLOPs, inference latency, training time per epoch, and peak GPU memory. For CASNet, the reported detector consists of the original visual encoder, the amplified visual encoder, and the MLP classifier. During downstream detector training, both visual encoders are frozen and only the MLP classifier is optimized. For a conservative and fair comparison with methods that recompute backbone features during training, we report the online-equivalent training time, where the frozen features are recomputed once per epoch before classifier optimization. In practice, because the two visual encoders are frozen, CASNet can cache the extracted features once; this reduces the classifier training cost to 5.01 seconds per epoch after a one-time feature extraction cost of 6682.08 seconds.

Results. Table 8 shows that CASNet has a larger total parameter count and higher GFLOPs than single-backbone detectors, which is expected because it uses both the original and amplified visual encoders to extract the alignment-shift increment. However, the additional cost mainly comes from frozen encoders rather than trainable modules: CASNet only optimizes 0.46M classifier parameters during downstream detector training, which is substantially fewer than fully trainable CNN-based detectors and VIB-Net. Despite using two visual encoders, CASNet achieves an online-equivalent training time of 6687.09 seconds per epoch, lower than all compared methods in the table. Its inference latency is 20.62 ms per image, which is higher than single-backbone UnivFD and VIB-Net but still lower than CLIPping. The main cost of CASNet is memory usage, where the dual-encoder design leads to a peak memory of 3698.50 MB. Overall, this experiment shows that CASNet trades moderate additional frozen-model computation and memory for a lightweight trainable detector and competitive training efficiency.

C Robustness to Image Post-processing

Purpose. Images shared in real scenarios are often modified by common post-processing operations such as resizing, compression, blur, noise, and color adjustment. Since many AIGI detectors rely heavily on low-level visual artifacts, these operations may weaken or remove the traces used for detection. We therefore conduct a robustness evaluation to examine whether CASNet can maintain reliable detection performance under such image degradations.

Table 9: Robustness comparison under common image post-processing operations. All results are reported in accuracy (%). The average column is computed over all 15 perturbation settings.

Method	Scale			Noise			Color		
	$0.25\times$	$0.5\times$	$0.75\times$	$\sigma = 5$	$\sigma = 10$	$\sigma = 20$	Bright. 0.4	Contrast 0.4	Hue 0.1
UnivFD	65.91	68.14	70.20	58.67	58.29	58.33	63.63	63.17	63.92
Effort	63.27	66.14	70.73	72.14	69.36	64.14	73.82	73.64	74.82
VIB-Net	62.04	64.85	65.81	71.54	69.92	64.85	72.27	71.96	73.12
CASNet	**70.08**	**75.12**	**78.46**	**80.62**	**76.73**	**69.00**	**85.77**	**85.58**	**86.04**

Method	JPEG			Blur			Avg.
	Q=90	Q=80	Q=70	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	
UnivFD	66.81	66.85	67.38	69.85	60.38	59.77	64.09
Effort	70.00	69.00	66.00	69.00	63.00	60.00	68.34
VIB-Net	72.62	71.38	73.19	71.77	65.42	62.77	68.90
CASNet	**76.42**	**74.85**	**75.00**	**75.81**	**70.27**	**66.92**	**76.44**

Protocol. We evaluate each detector without retraining on perturbed test images. The perturbation suite contains five types of post-processing operations: down-scaling, additive Gaussian noise, color jitter, JPEG compression, and Gaussian blur. For down-scaling, images are resized to $0.25\times$, $0.5\times$, or $0.75\times$ of the original size and then resized back. For noise, we add pixel-space Gaussian noise with $\sigma \in \{5, 10, 20\}$. For color changes, we apply brightness adjustment, contrast adjustment, and hue shift. For compression and blur, we evaluate JPEG quality factors of 90, 80, and 70, and Gaussian blur with $\sigma \in \{1, 2, 3\}$. We report accuracy averaged over the evaluated test sources.

Results. Table 9 shows that CASNet is consistently robust to common image post-processing operations. It achieves the best accuracy in all 15 perturbation settings and obtains the highest average accuracy of 76.44%, compared with 64.09% for UnivFD, 68.34% for Effort, and 68.90% for VIB-Net. Compared with the strongest baseline in each perturbation group, CASNet improves the average accuracy by 6.47, 6.68, 11.70, 3.03, and 4.35 percentage points under scale, noise, color, JPEG compression, and blur perturbations, respectively. Overall, this experiment demonstrates that the proposed alignment-shift representation provides a stable cue for generated-image detection under practical image degradations.

D Limitations

CASNet relies on pretrained vision–language backbones and paired text–real–generated triplets to amplify cross-modal alignment shift. Although our experiments show that CAS transfers across multiple backbones and generators, the strength of the alignment-shift signal may vary for domains or generation pipelines whose image–text alignment patterns differ substantially from those used during amplification. In particular, the paired construction of Sec. 3.1, in which captions are reused as generation prompts and the most similar of ten candidates is kept, can magnify the measured shift, and CAS is strongest and most consistent for text-conditioned generation; for class-conditioned or unconditional generators, as well as image-to-image generators, the shift is expected to be weaker and backbone-dependent, so that CASNet relies more on the conventional visual evidence in $\tilde{z}^{\mathrm{amp}}$ than on the alignment-shift increment. This is consistent with the weaker columns in Tables 1–3: non-text-conditioned or image-to-image generators (ADM, GauGAN, CycleGAN), and heavily stylized commercial generators (Midjourney, Hourglass) that lie far from the natural-photo, descriptive-caption Flickr domain used for amplification. We also expect CASNet to be weaker on deliberately prompt-misaligned images, edits and composites, adversarially perturbed images, and camera-recaptured content.

CASNet also compares representations from the original and amplified visual encoders, which introduces additional computation and memory compared with single-encoder detectors. This

overhead is mainly caused by frozen encoders and can be reduced by feature caching during detector training, but developing more lightweight CAS extraction strategies remains an interesting direction for future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions: identifying Cross-modal Alignment Shift (CAS), proposing CASNet, and evaluating it across diverse generators. These claims are supported by the empirical analysis in Section 3, the method in Section 4, and the results in Section 5 and Appendix C.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations section in Appendix D. It discusses dependence on pretrained vision–language backbones and paired text–real–generated triplets, as well as the additional computation and memory caused by the dual-encoder design.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theoretical results, theorems, or proofs. The mathematical expressions are used to define the empirical measurements, losses, and representations of the proposed method.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the paired data construction, model architecture, training objectives, backbone choice, optimizer, hyperparameters, datasets, baselines, and evaluation protocols in Sections 3–5 and the appendices. These details provide the information needed to reproduce the main empirical claims.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and model weights are publicly available at `https://github.com/oceanzhf/CASNet`, together with the datasets, generated-data construction procedure, model settings, and implementation details specified in Section 5.1 and the appendices.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5.1 specifies the training and test datasets, generation sources, backbone, frozen layers, optimizer settings, learning rates, batch sizes, epochs, and hyperparameters. The appendices further describe the additional CAS analysis, computational profiling, and robustness evaluation protocols.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: The paper supports its main empirical claims with extensive evaluations across diverse datasets, generators, backbone models, and image perturbations, reported in both the main paper and the appendix. It further includes ablation studies and auxiliary analyses to verify that the observed improvements are consistent rather than isolated to a single setting. While the paper does not focus on formal significance testing, the breadth and consistency of the empirical results provide sufficient support for the conclusions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 5.1 states the GPU type used in the experiments, and Appendix B reports parameters, GFLOPs, inference latency, per-epoch training time, and GPU memory usage. These measurements characterize the compute requirements of CASNet relative to representative baselines.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is intended to improve detection of AI-generated images and uses standard public datasets, generated images, and pretrained models. It does not involve human-subject experiments or release image-generation systems that would directly increase generative misuse risk.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The introduction discusses the positive motivation of mitigating misinformation and improving trust in digital media.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk assets such as pretrained language models, image generators, scraped datasets, or tools designed to generate harmful content. The proposed method is a detector, and no special release safeguard is described.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the datasets, models, and baseline methods used in the experiments, including Flickr30k, GenImage, ForenSynths, AIGI-Holmes, Community Forensics, and pretrained vision–language backbones.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release a new dataset, benchmark, model checkpoint, or code package as a documented asset in the current submission. The generated paired samples are used for training and analysis, but no new asset release is described.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or new research with human subjects. All experiments use existing datasets, generated images, and automated evaluation protocols.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research, so IRB approval or equivalent review is not applicable. No study participants are recruited or exposed to research risk.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Section 3.1 describes the use of ChatGPT to merge the five Flickr30k captions associated with each image into a single description for constructing paired text–real–generated triplets. LLM usage is limited to this data-construction step and is not part of the CASNet detector at inference time.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.