Scaling Up AI-Generated Image Detection with Generator-Aware Prototypes

Supplementary Material

We organize the supplementary in the following way:

- Sec. A: Detailed information of the introduced series of dataset build in previous section.
- Sec. B: Details of the 6 selected benchmarks.
- Sec. C: Introduction to the selected baselines.
- Sec. D: Detailed results of each test subsets.
- Sec. E: Implementation details of the proposed method.
- Sec. F: Additional derivation and analysis.
- Sec. G: Future perspectives towards reasoning-driven and embodied forensics.

A. Settings of the toy dataset

In Sec.3, we introduce a series of datasets, comprising generated images from different numbers of generators. In this section we will give the constructing procedure of these datasets.

First, there are 8 generators in GenImage [52] dataset, in each generator subset, there are 1000 types of object mirroring the 1000 categories of ImageNet-1k [13]. We select generator subset composed of n_g generators based on Tab. 1. For each of the 1000 categories, we randomly sample n_s images in each generator subset. To pair real images, we randomly sample $n_s \times n_g$ images from the original ImageNet dataset. n_s is determined via $1000 \times n_s \times n_g = 8000$.

For the last dataset, which consist of thousands of generators, we leverage Community-Forensics training dataset [33] for collecting, which is the same as our training dataset. we randomly sample 2 images from each generator in it, which consist of about 9000 generated images. Then we randomly sample 8,000 real images as before to construct the last dataset in the series.

group	n_g	Generator(s)
1	1	SDv1.4
2	2	SDv1.4, BigGAN
3	4	SDv1.4, BigGAN, VQDM, Glide
4	8	All GenImage

Table 1. Generators used to build our datasets.

B. Benchmarks

We select 6 benchmarks to represent most existed generative models to evaluate the methods. Though some subsets have same or similar architecture, their training condition, sampling strategy and semantic content are not quite the same. Thus we preserve all subsets that have same name to simulate a variety of generated images.

Forensic Synthetic [47] contains a series of CNN-generated images, we select its GAN variants, including ProGAN [23], StyleGAN [24], StyleGAN2, CycleGAN [51], StarGAN [11], GauGAN [34], BigGAN [8], and Deepfake [37] for forgery face.

UFD [32] datasets expand the dataset above by introducing several early diffusion models and commercial APIs, including latent diffusion model [36], Glide [31] and Guided [14] diffusion model.

GenImage [52] provide a dataset trained on ImageNet-1k. It has 8 generative models in both GANs, Diffusion Models and Commercial APIs, including BigGAN [8], VQDM [16], Stable Diffusions, Wukong [2], ADM [14] and Midjourney [1].

SynthBuster [6] provide an aligned dataset, where real images and generated images are all in PNG format, which makes it challenging for AIGI detectors that leverage format shortcut. Moreover the generated images are also from popular latent diffusions, including DALL-E, Stable Diffusions, Firefly[3] and Glide.

Chameleon [49] provides a sanity check for AI-generated image detection. They build a high quality dataset where generated images are source from internet and some unknown source. All images in this benchmark are said to be indistinguishable by human. Since all images are gather from the unknown source, there's only one subset in this benchmark.

Community Forensics Evaluation Set [33] is build to evaluate the model's ability to generalize to unseen generators that trained in Community Forensics dataset. This evaluation dataset is also the most up-to-date dataset, containing generators like Deci Diffusion V2 [46], GALIP [44], KandinskyV2.2 [38], Kvikontent [25], LCM-LoRA-SDv1.5, LCM-LoRA-SDXL, LCM-LoRA-SSD1B [28], Stable Cascade [35], DF-GAN [43], and HDiT [12].

Above all, we have 55 subsets for testing. Given the large scale of our training data, the training domain overlaps with several previously constructed datasets. Consequently, our evaluation comprises 29 completely unseen generator subsets, the rest, even though seen in training set, still have a different generated condition. Sample images from the test set are shown in Fig. 1.

Metrics. Following prior works [32, 40, 47], we compute a threshold-free metric, mean average precision (AP), and a threshold-based metric, binary accuracy (Acc). When

Figure 1. Examples of test subsets, we visualize some in-domain datasets with our training set along with some out-of-distribution sets.

computing accuracy, the threshold was set to 0.5.

C. Baselines

In this section, we will give a brief introduction to the baselines for comparison.

C.1. GAN-Generalized

CNNDetection [47] uses a ResNet-50 as a classifier with data augmentation to detect CNN-generated images. **NPR** [40] rethink up-sampling operation in most generative architecture and detect them via a interpolation pattern. **UniFD** [32] leverage the image encoder of CLIP for feature extraction, it takes image embeddings for classification with simple KNN or linear layer. **SAFE**[26] extracts high frequency band as artifact with various data augmentation to build a CNN classifier. **AIDE** uses a hybrid feature of both CLIP image embedding , high and low frequency patches via DCT scoring, and concatenate them for final decision. These baselines are all trained on GAN generated images provided by the training set of CNNDetection [47].

C.2. Diffusion-Generalized

In the diffusion era [36], more method aim to detect generated images with a more realistic diffusion generated images. **DRCT** [9] construct a diffusion reconstruct dataset and used it to train a classifier with classification objective and contrastive learning objective. We use its Conv-B variant trained on SDv2.1. **Co-SPY** [10] also uses a hybrid feature for classification, but a combination of CLIP embedding and VAE-reconstruct residual. **B-Free** [17] proposed a paradigm that generated images should come from inpainting models rather than unconditional generators or T2Is to prevent semantic bias. It trained an end-to-end classifier with proposed dataset.

C.3. Scaling-Ups

Scaling up detectors use a training dataset from a more diverse source. To solve the sanity check for AIGI detection, AIDE [49] provide another checkpoint that trained a classifier with all images in GenImage, including both GANs and Diffusions. D3 [50] proposed a dual feature extraction branch, a original image and a patch-shuffled image to learn comprehensive traces, it uses a training dataset consist of GenImage and CNNDetection. Community Forensics construct a large dataset with thousands of generators and use it to train classifier with a standard ViT.

D. Detailed results of experiments

The following tables show detailed results for the selected 6 benchmarks. All baselines are evaluated with their provided checkpoints. Tab. 2 shows results of ForenSynths, Tab. 3 shows results of UFD, Tab. 4 shows results of GenImage, Tab. 5 shows results of SynthBuster, Tab. 6 and Tab. 7 shows results of Community Forensics evaluation set.

E. Implementation Details

To validate the reproducibility, we present our implementation details in this section.

Network Design The image encoder we use is CLIP:ViT-L with a patch size of 14×14. The MLP is composed of two hidden layer with dimensions $1024 \rightarrow 128 \rightarrow 1$. In the second stage, we discard the second layer to make the MLP $1024 \rightarrow 128$, then we apply LoRA [18] in `q_proj, k_proj, v_proj` of image encoder with LoRA parameter $r = 16, \alpha = 32$, dropout $= 0.1$.

Training For image in training set, we apply RandomCrop to 224×224, and to those image whose resolution is smaller than this, we apply 0-padding to make it enough to crop. We automatically separate 5% samples of dataset to form a validation set. We use AdamW [27] optimizer with

a learning rate of 10^{-4} and a weight decay of 0.01 in the two stages. In the first stage, training last for 20 epochs. In the second stage. we train until the validation accuracy reaches 99.9% or do not improve for 3 epochs to prevent performance degradation.

Testing For images in testing sets, we apply CenterCrop to 224×224, which is the same as most baselines, except for **B-Free**, who uses a resolution of 504×504 and **Co-SPY** uses a 384×384, to best excel their performance, we did not modify their model but directly report their performance on original resolution.

F. Additional Derivation and Analysis

In this section, we provide supplementary materials to substantiate the claims made in the main text. Specifically, we present a rigorous mathematical derivation modeling the data heterogeneity in scaling-up settings (Sec. F.1) and offer qualitative visualizations of the attention maps to demonstrate how our method alters the encoder's focus (Sec. F.2).

F.1. Detailed Derivation of Theory

Recall that in Sec. 3 of the main paper, we identified the "Benefit then Conflict" dilemma, attributing it to the severe data-level heterogeneity that arises when aggregating multiple generators. To theoretically quantify this phenomenon, we model the distribution of real and generated images and analyze the behavior of feature variance.

Why GMM? How to calculate variance in GMM?

We adopt the Gaussian Mixture Model (GMM) as a proxy for the real-world data distribution. This choice is motivated by the structure of commonly used datasets like ImageNet, which consist of distinct categories. Specifically, we model the image features within each category as a multivariate Gaussian distribution. While this is a simplified assumption, it remains theoretically robust and yields conclusions consistent with more complex scenarios. The total variance of the generated distribution is derived as follows:

$$\Sigma_{gen} = \mathbb{E}_G \left[\text{Var}(X|G)\right] + \text{Var}_G \left[\mathbb{E}(X|G)\right], \tag{1}$$

For the i-th generator, its variance is computed by:

$$\begin{aligned} &Var(X|G=i) = \Sigma^i_{gen} \\ &= \sum_{j=1}^{M} \pi_{i,j}\Sigma_{i,j} + \sum_{j=1}^{M} \pi_{i,j}(\mu_{i,j}-\mu_i)(\mu_{i,j}-\mu_i)^T \end{aligned} \tag{2}$$

Thus, the total variance can be expressed as:

$$\begin{aligned} \Sigma_{gen} &= \sum_{i}^{G} w_i \sum_{j}^{M} \pi_{i,j}\Sigma_{i,j} \\ &+ \sum_{i}^{G} w_i \sum_{j}^{M} \pi_{i,j}(\mu_{i,j}-\bar{\mu})(\mu_{i,j}-\bar{\mu})^T \\ &+ \sum_{i}^{G} w_i (\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^T \\ &= \underbrace{\mathbb{E}_M \left[\text{Var}(X|G_i, M)|G_i\right] + \text{Var}_M \left[\mathbb{E}(X|G_i, M)|G_i\right]}_{\text{generator fitting variance}} \\ &+ \underbrace{\text{Var}_G \left[\mathbb{E}(X|G)\right]}_{\text{cross generator variance}}, \end{aligned} \tag{3}$$

where $\mu_i = \sum_j^M \pi_{i,j}\mu_{i,j}$, and $\bar{\mu} = \sum_i^G \mu_i$ denotes the mean within a generator and the global mean across the entire dataset, respectively.

How do prototypes limit variance growth?

Having established that variance increases with generator diversity, we now show how our prototype mapping strategy mitigates this issue. In prototype mapping, the feature embedding is reorganized into a linear combination of prototypes; this operation effectively sets an upper bound on the embedding's variance.

We have $\tilde{f} = \sum w_i v_i$ where w_i is the attention score and v_i is a prototype vector. Let F be the random variable of this reorganized embedding. Its variance is given by:

$$Var(F) = \mathbb{E}\left[||F - \mathbb{E}(F)||^2\right], \tag{4}$$

where $Var(\cdot)$ here denotes its trace in the covariance matrix, i.e., $tr(\Sigma)$. To proceed, we introduce an independent identically distributed random variable F' to reformulate this term as:

$$\begin{aligned} Var(F) &= \frac{1}{2}\mathbb{E}\left[||F - F'||^2\right] \\ &= \frac{1}{2}\sum w_i w_j ||v_i - v_j||^2, \end{aligned} \tag{5}$$

Assume $D = \max_{v \in P} ||v_i - v_j||$. Since all the prototypes are determined and fixed, this maximum distance is a constant. Due to the fact that $w_i, w_j \in [0,1]$ and $\sum w_i = 1$, this variance term can be bounded by:

$$Var(F) \leq \frac{1}{4}D^2. \tag{6}$$

This derivation confirms that regardless of the number of source generators, the variance of the features mapped to the prototype space remains bounded by the geometry of the prototype set.

F.2. Visualization of Attention Maps

To provide interpretability for our model's performance, we move beyond theoretical bounds to empirical visualization. We compare the attention maps from both the original CLIP image encoder and the encoder fine-tuned by our proposed GAPL.

These attention maps represent the attention scores between the [CLS] token and all spatial patch tokens. The image encoder consists of 24 ViT blocks; in our visualization, we sample 8 blocks, with block indices ascending from left to right (shallow to deep). For visual clarity, bicubic interpolation is applied to the raw attention maps. As shown in Fig. 2 and Fig. 3, the comparisons demonstrate two key findings: (1) Our encoder preserves rich semantic information in the shallow layers, maintaining the generalization capability of the pre-trained model; (2) In deeper layers, compared to the original CLIP, our model effectively leverages more patches to form a more comprehensive artifact pattern. Visually, this manifests as more pronounced and focused bright spots in the deeper layers, indicating that GAPL has successfully learned to attend to generator-specific traces.

G. Future Perspectives

While our proposed GAPL framework achieves state-of-the-art performance by managing data heterogeneity through generator-aware prototypes, we recognize that the arms race between generation and detection is evolving. As generative models (e.g., Flux [7], Midjourney v6 [30], Nano Banana Pro [15]) increasingly mitigate low-level statistical artifacts, relying solely on texture or frequency cues may face diminishing returns. We envision that the next generation of AIGI detectors must integrate higher-level cognitive capabilities. Specifically, we identify three promising avenues where our scaling-up principles can intersect with broader AI domains:

From Artifact Detection to Visual Reasoning. Current forensic methods predominantly focus on signal-level anomalies. However, highly realistic AI-generated images often retain subtle *semantic* or *logical* inconsistencies (e.g., impossible shadows, mismatched reflections, or counting errors) that are invisible to standard classifiers. Future work should explore integrating **Visual Reasoning** frameworks [20, 22, 39, 41] into the detection pipeline. By leveraging neuro-symbolic approaches or chain-of-thought reasoning, a detector could move beyond binary classification to provide interpretable evidence based on compositional logic, effectively identifying "why" an image violates reality even when pixel statistics appear perfect.

Incorporating Spatial Intelligence and Physical Constraints. A persistent weakness in current generative models is their frequent violation of 3D geometry and physical laws. While GAPL effectively learns 2D prototypes, it does not explicitly model the underlying 3D structure of the scene. Incorporating spatial intelligence [21, 45, 48] could serve as a powerful orthogonal check. Models equipped with 3D-aware representations or spatial reasoning capabilities can detect geometric implausibilities—such as "Escher-like" structures or inconsistent depth cues—that purely 2D generative models fail to resolve. This aligns with our observation of "Benefit then Conflict," suggesting that physical laws could serve as the ultimate invariant feature across diverse generators.

Trustworthy Perception in Embodied AI. Finally, the utility of AIGI detection extends beyond digital media forensics into the physical world. As we deploy autonomous agents, ensuring trustworthy perception is critical. Embodied agents operating in the real world must distinguish between authentic sensory inputs and potentially manipulated streams (e.g., adversarial projections or deepfakes in video feeds) [4, 19]. Furthermore, robust detection models like GAPL can act as quality filters for Embodied AI [5, 22, 29, 42], particularly in Sim-to-Real pipelines where agents are trained on synthetic data. By rigorously curating high-fidelity synthetic data that adheres to physical realism, we can ensure that embodied agents develop robust representations that generalize better to physical reality.

Method	ProGAN		StyleGAN		StyleGAN2		StarGAN		GauGAN		CycleGAN		BigGAN		Deepfake		Mean	
	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP
CNNDet [47]	100.0	100.0	74.3	98.4	76.3	92.1	81.1	95.3	80.1	98.0	81.1	96.4	59.5	88.2	51.0	66.1	75.3 ± 13.9	91.8 ± 10.3
NPR [40]	99.8	99.9	97.3	99.9	99.5	100	99.7	100	79.1	80.2	94.5	97.6	83.6	84.7	73.7	72.5	90.1 ± 9.9	91.8 ± 10.3
UniFD [32]	99.8	100.0	85.4	97.5	74.5	97.6	96.0	99.5	99.4	100.0	98.3	99.8	94.7	99.2	67.5	80.3	89.4 ± 11.6	96.7 ± 6.3
SAFE [26]	99.8	100.0	97.6	99.8	98.7	100.0	99.8	100.0	92.2	96.9	99.1	99.8	89.5	95.1	93.2	97.2	96.2 ± 3.7	98.6 ± 1.8
AIDE [49]	96.3	99.8	97.2	99.7	98.1	99.8	98.4	99.9	76.9	96.0	95.6	99.4	78.3	96.7	54.2	68.7	86.8 ± 14.9	95.0 ± 10.0
DRCT [9]	50.2	50.2	49.1	49.1	49.3	45.3	38.3	38.7	49.9	43.2	49.3	44.6	49.7	47.8	64.7	78.4	50.1 ± 6.7	49.6 ± 11.4
Co-SPY [10]	74.7	78.1	63.7	69.8	59.7	62.9	62.1	94.3	69.6	83.4	58.5	55.8	71.6	83.9	64.9	78.7	65.6 ± 5.4	75.9 ± 11.6
B-Free [17]	95.6	99.3	74.8	93.6	71.1	89.4	81.1	93.5	96.0	99.8	65.4	90.5	91.5	98.9	73.7	89.2	81.2 ± 11.1	94.3 ± 4.2
AIDE† [49]	92.7	98.8	88.6	95.1	93.6	98.6	88.0	99.0	87.7	98.9	92.0	98.9	84.9	97.7	54.8	63.2	85.3 ± 11.8	93.8 ± 11.6
D3 [50]	99.8	100.0	93.5	99.1	95.8	99.5	93.5	98.7	98.7	100.0	95.9	99.5	99.6	100.0	67.5	87.0	93.0 ± 9.9	98.0 ± 4.2
CommForen [33]	92.8	99.8	93.1	99.3	92.7	99.5	98.8	99.9	98.8	99.9	95.5	99.8	99.6	100.0	66.8	88.9	92.2 ± 10.0	98.2 ± 3.6
GAPL(Ours)	99.9	100.0	98.1	100.0	99.5	100.0	97.0	99.9	99.6	100.0	98.2	99.9	98.6	100.0	88.2	96.8	97.2 ± 3.6	99.5 ± 1.1

Table 2. Detaild results on the benchmark ForenSynth [47]. we only select its GAN variant. AIDE† denotes its scaling up checkpoint trained on 8 generators.

Method	DALL-E		Glide-50-27		Glide-100-10		Glide-100-27		Guided		LDM-100		LDM-200		LDM-200-cfg		Mean	
	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP
CNNDet [47]	52.9	68.4	55.7	78.0	54.3	74.3	53.3	73.7	52.8	68.1	52.0	68.7	51.5	68.2	52.2	69.6	53.1 ± 1.3	71.1 ±3.5
NPR [40]	90.3	97.4	97.2	99.2	97.6	99.2	96.9	99.1	87.1	92.6	97.4	99.1	97.6	99.2	97.4	99.0	95.2 ± 3.8	98.1 ± 2.2
UniFD [32]	87.5	97.7	79.2	96.0	78.0	95.5	78.7	95.8	70.0	88.3	95.2	99.3	94.5	99.4	74.2	93.2	82.2 ± 8.7	95.7± 3.4
SAFE [26]	97.5	99.7	96.6	99.2	97.3	99.4	95.8	98.9	82.4	95.8	98.8	100.0	98.8	100.0	98.7	99.9	95.7 ±5.1	99.1 ± 1.3
AIDE [49]	89.7	99.6	89.9	99.8	89.8	99.6	89.9	99.8	94.2	98.9	90.1	99.9	90.1	99.9	90.1	99.9	90.5 ±1.4	99.7 ± 0.3
DRCT [9]	55.6	57.8	56.2	62.6	61.0	70.4	56.2	62.7	62.6	89.3	88.8	96.6	88.9	96.7	90.3	97.5	70.0 ± 15.2	79.2 ±16.3
Co-SPY [10]	81.8	87.2	69.0	74.6	76.7	81.7	73.5	78.2	62.7	87.4	82.7	86.9	83.1	87.5	85.2	91.0	76.8 ±7.4	84.3 ±5.2
B-Free [17]	93.2	97.9	78.6	90.2	81.8	91.9	77.9	89.4	74.6	93.6	97.2	99.9	97.1	99.8	96.9	99.8	87.1 ±9.2	95.3 ± 4.2
AIDE† [49]	98.7	99.9	99.0	100.0	99.1	100.0	99.0	100.0	95.5	99.9	99.1	100.0	99.1	100.0	99.1	100.0	98.6 ± 1.1	100.0 ± 0
D3 [50]	94.0	98.7	95.8	99.3	95.8	99.4	95.7	99.5	95.9	99.6	96.1	99.7	95.8	99.7	89.3	96.5	94.8 ± 2.2	99.0 ± 1.0
CommForen [33]	98.4	99.9	97.1	99.6	98.2	99.8	96.8	99.6	66.1	76.6	98.5	99.9	98.7	99.9	98.5	99.9	94.0 ± 10.5	96.9 7.6
GAPL(Ours)	97.8	100.0	97.7	99.9	97.8	100	97.7	100.0	93.4	98.6	97.8	100.0	97.9	100.0	97.9	100	97.2 ± 1.5	99.8 ± 0.5

Table 3. Detaild results on the benchmark UFD. We refer this set to the diffusion part that [32] added.

Method	VQDM		SDv1.4		BigGAN		Wukong		SDv1.5		Glide		Midjourney		ADM		Mean	
	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP
CNNDet [47]	52.1	68.8	50.5	59.8	51.2	78.6	50.8	59.3	50.7	60.6	52.3	67.6	51.3	64.7	51.5	65.6	51.3 ± 0.6	65.6 ± 5.9
NPR [40]	89.6	94.1	92.1	95.0	70.1	80.3	88.0	93.0	91.7	95.4	93.2	96.5	79.7	86.5	85.7	91.7	86.2 ± 7.3	91.6 ± 5.1
UniFD [32]	84.0	96.3	64.1	87.4	89.8	98.5	71.8	91.2	64.4	86.0	61.6	84.6	57.2	75.5	67.1	86.8	70.0 ± 10.6	88.3 ± 6.7
SAFE [26]	96.1	99.6	99.4	100.0	97.6	99.8	98.1	99.8	98.8	99.9	97.2	99.6	95.5	99.5	81.5	96.4	95.5 ± 5.4	99.3 ± 1.1
AIDE [49]	98.4	99.9	98.3	99.9	98.9	99.9	97.3	99.8	98.1	99.8	98.6	99.9	86.4	96.6	96.8	99.5	96.6 ± 3.9	99.4 ± 1.1
DRCT [9]	61.3	85.4	97.9	99.8	52.3	86.3	95.9	99.6	97.9	99.8	60.1	91.2	98.2	99.9	60.0	90.2	78.0 ± 20.0	94.0 ± 6.0
Co-SPY [10]	72.2	93.9	93.1	99.2	64.7	90.8	91.2	98.8	93.0	99.0	81.4	96.5	70.0	91.0	58.8	84.3	78.0 ± 12.7	94.2 ± 4.9
B-Free [17]	88.2	98.3	99.4	100.0	76.5	96.8	99.4	100.0	99.3	100.0	69.7	93.8	94.1	99.3	72.5	93.7	87.4 ± 11.9	97.7 ± 2.5
AIDE† [49]	99.9	100.0	99.8	100.0	99.9	100.0	99.6	100.0	99.8	100.0	99.8	100.0	99.0	100.0	99.5	100.0	99.7 ± 0.2	100.0± 0
D3 [50]	98.3	99.9	98.1	99.8	97.1	99.7	97.7	99.8	97.7	99.8	98.3	99.8	79.7	95.8	96.8	99.6	95.5 ± 6.0	99.3 ± 1.3
CommForen [33]	90.9	99.7	90.6	99.2	71.5	81.6	91.0	99.1	90.5	99.2	89.4	98.4	77.5	86.8	72.9	83.5	84.3 ± 8.1	93.4 ± 7.5
GAPL(Ours)	98.2	100	98.2	100	97.9	99.8	98.1	99.9	98.0	99.9	98.1	99.9	90.3	97.6	95.0	99.2	96.7± 2.6	99.6 ± 0.7

Table 4. Detaild results on the benchmark GenImage.

Method	Real	DALLE-2		Firefly		SDv1.4		SDXL		DALLE-3		Glide		MJ-v5		SDv1.3		SDv2		Mean	
	Acc	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP
CNNDet [47]	96.5	7.0	64.8	14.4	74.9	1.5	49.9	5.2	59.0	0.1	36.2	1.8	46.9	1.2	44.8	1.3	49.3	4.8	96.5	50.3 ± 2.1	54.0 ± 17.9
NPR [40]	5.6	98.4	49.6	5.2	33.3	94.3	50.6	100	54.7	15.5	31.0	91.6	50.8	80.5	46.1	94.3	50.0	70.4	42.7	38.9 ±17.1	45.4±7.7
UniFD [32]	93.5	77.0	95.2	86.0	97.4	45.6	85.3	41.4	83.5	0.7	41.2	11.5	63.5	10.4	61.0	44.8	85.3	58.6	89.9	67.6 ±14.0	78.0 ± 17.6
SAFE [26]	15.5	92.0	42.2	4.3	31.0	91.8	55.9	87.4	37.1	44.5	34.7	58.1	36.0	97.0	54.8	91.0	53.9	99.5	56.5	44.7±15.1	44.7 ± 9.8
AIDE [49]	66.7	45.7	57.0	0.0	30.8	95.1	94.8	98.2	94.3	2.8	35.9	96.7	95.7	75.3	80.3	95.8	95.3	95.6	93.3	67.0 ± 19.3	75.3 ±25.3
DRCT [9]	96.1	4.1	53.6	11.4	60.7	88.2	97.9	89.6	98.2	35.6	80.8	14.1	72.9	99.4	99.9	89.6	98.2	99.9	100.0	77.6 ± 19.6	84.7± 17.3
Co-SPY [10]	97.6	48.5	90.8	43.6	87.5	74.7	96.4	44.1	87.9	73.7	96.5	80.6	97.8	35.2	85.0	74.5	96.7	40.0	85.4	77.4 ± 8.6	91.6 ± 5.0
B-Free [17]	98.5	88.5	98.9	99.2	99.9	99.8	99.9	100.0	100.0	93.1	99.4	42.7	91.8	98.9	99.9	100.0	100.0	99.5	99.9	94.9± 8.8	98.9 ± 2.5
AIDE† [49]	32.8	27.3	35.8	0.3	31.5	99.0	64.4	98.6	64.8	34.8	43.4	97.1	66.8	98.6	75.0	99.7	65.8	96.7	57.0	52.7 ± 18.8	56.0 ± 14.5
D3 [50]	82.0	87.7	92.6	92.0	95.3	96.5	97.9	89.4	93.6	28.6	61.3	90.6	95.3	62.6	81.2	96.4	97.7	82.6	90.8	81.3 ± 10.4	89.5 ± 11.0
CommForen [33]	84.6	83.9	91.6	93.1	96.6	94.7	97.4	98.5	98.3	76.1	88.4	90.0	96.0	80.6	91.4	95.7	97.7	92.9	95.3	87.1 ±3.6	94.8 ± 3.3
GAPL(Ours)	90.0	94.0	97.5	94.3	98.2	98.1	99.3	99.8	99.8	60.1	85.3	97.6	99.2	92.0	97.4	98.2	99.3	96.4	98.7	91.1 ± 5.8	97.2 ±4.3

Table 5. Detailed results on the benchmark SynthBuster [6]. In this benchmark, all generated images are pair with exactly the same real images. Thus we report the real images accuracy and each generator's fake accuracy independently. We pair the metrics of real and each generators to get the final mean metrics .

Method	DFGAN		MJv6		Kandinsky		SDcas.		MJv5		Firefly2		Firefly3		GALIP		LCM-lora-sdxl		Hourglass		Kvikontent	
	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP
NPR [40]	96.6	100.0	98.6	99.0	57.7	56.1	57.4	56.4	96.7	98.3	49.2	45.0	49.6	52.4	51.7	48.0	39.9	41.6	88.6	91.6	58.6	56.3
SAFE [26]	50.3	44.8	49.9	45.0	49.6	53.2	49.9	51.0	49.9	45.1	50.0	51.3	50.1	51.2	49.5	53.9	49.5	52.3	50.1	45.2	50.2	51.5
AIDE [49]	49.9	43.9	49.8	44.7	49.9	50.1	50.2	52.6	49.9	44.4	50.2	46.2	50.2	46.8	49.7	53.0	49.7	46.1	49.8	46.1	50.1	54.1
DRCT [9]	49.9	45.9	49.6	47.0	49.6	53.1	49.4	52.7	49.7	50.5	49.6	49.1	49.9	47.0	50.2	56.6	50.0	49.0	50.0	46.9	49.6	52.2
Co-SPY [10]	50.0	84.5	70.1	86.5	68.4	74.8	71.0	76.7	63.0	80.8	68.6	86.4	86.7	95.8	38.4	34.1	43.1	38.8	52.8	59.1	74.4	84.8
B-Free [17]	92.2	96.5	76.7	84.9	85.7	99.4	86.1	99.3	86.2	97.2	82.9	92.7	81.8	91.7	81.8	91.9	85.0	97.6	57.6	65.6	86.6	99.5
AIDE† [49]	49.9	56.4	49.7	46.9	50.4	53.3	50.8	53.8	50.2	47.5	49.7	50.7	50.0	47.8	50.4	55.9	50.4	51.8	50.0	44.7	50.3	48.1
D3 [50]	98.4	100.0	69.5	81.0	67.9	84.5	71.5	90.0	76.7	87.0	80.0	90.8	78.6	88.7	70.8	90.8	63.0	77.7	68.3	83.1	74.9	95.7
GAPL(Ours)	99.0	100	87.1	97.3	93.0	99.1	93.3	99.6	88.6	99.3	87.3	96.2	85.1	93.7	90.6	98.1	88.2	95.8	78.5	89.4	94.5	99.8

Table 6. Detailed results on the benchmark Community forensic evaluation. This is the first part. Note that the results of CNNDet, UniFD and CommForen are directly cited from the original paper [33] and dataset repository, whose detailed results are not available.

Method	DALL-E 2		DALL-E 3		LCM-lora-sdv1.5		DeciDiff.		FLUX-dev		FLUX-schnell		IdeogramV2		IDeogramV1		Imagen3		LCM-lora-ssd1b		Mean	
	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP
NPR [40]	91.7	97.6	96.5	99.2	53.6	50.1	43.9	44.1	98.1	98.9	97.8	99.0	96.4	98.8	96.5	98.5	98.6	98.7	32.3	38.4	73.8 ± 24.0	74.7 ± 25.0
SAFE [26]	50.7	44.8	50.1	45.1	49.7	50.8	49.7	50.5	49.9	45.1	50.1	45.0	50.1	45.0	49.9	45.1	50.1	44.9	50.0	53.2	50.0 ± 0.3	48.3 ± 3.5
AIDE [49]	49.6	44.2	50.0	45.2	50.1	52.4	50.0	48.5	49.9	44.9	49.8	44.5	50.0	44.4	49.9	44.7	49.8	45.0	50.0	47.7	49.9 ± 0.2	47.1 ± 3.2
DRCT [9]	49.3	46.8	49.5	49.0	49.8	51.5	49.6	51.8	48.4	46.8	48.6	47.0	50.0	49.1	49.6	49.4	49.0	49.0	49.2	46.9	49.5 ± 0.4	49.4 ± 0.7
Co-SPY [10]	77.4	91.3	85.2	96.0	67.2	74.1	71.7	79.8	80.1	93.5	71.8	87.6	71.2	88.3	75.6	90.8	77.0	91.8	61.0	68.3	67.9 ± 12.4	79.2 ± 16.6
B-Free [17]	74.3	81.3	86.5	98.9	85.5	99.0	86.0	99.3	73.8	82.7	79.0	88.2	81.2	89.2	81.8	91.4	74.6	83.0	87.0	98.2	81.5 ± 7.1	91.8 ± 8.4
AIDE† [49]	49.3	47.4	49.9	45.1	50.6	48.2	50.4	50.7	49.9	40.9	50.1	41.7	50.0	42.9	50.0	43.7	50.1	44.4	50.0	51.8	50.1 ± 0.3	48.3 ± 4.4
D3 [50]	87.5	95.0	90.1	96.6	67.0	80.6	62.1	76.4	79.3	89.7	73.5	83.8	67.2	77.5	67.6	78.6	70.7	83.1	61.5	74.5	73.6 ± 9.3	86.0 7.1
GAPL(Ours)	88.4	97.5	88.9	99.6	93.9	99.7	92.3	99.1	87.3	98.9	88.0	98.2	88.0	98.2	87.7	98.7	87.0	95.9	90.4	97.6	89.4± 4.0	97.8± 2.4

Table 7. Detailed results on the benchmark Community forensic evaluation. This is the second part.

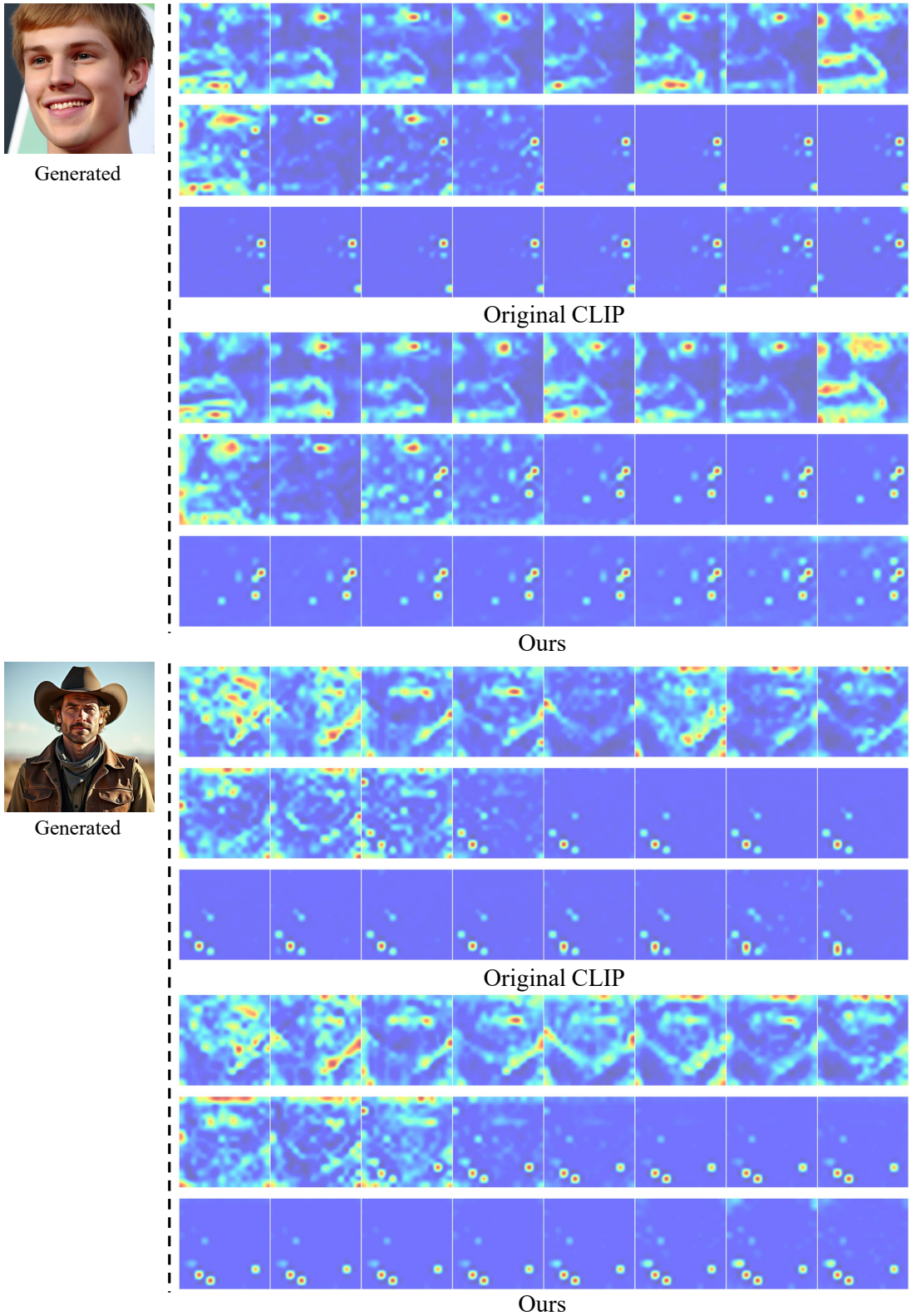

Figure 2. **Self-Attention map** between original CLIP backbone and our finetuned backbone. There are 24 ViT blocks in the image encoder, we plot 8 blocks in each row, with indices increasing from left to right. For clarity of visualization, we use bicubic interpolation between image patches. In the shallow layers, we preserve most semantic features, in deep layers, our attention includes a wider range compared to original CLIP.

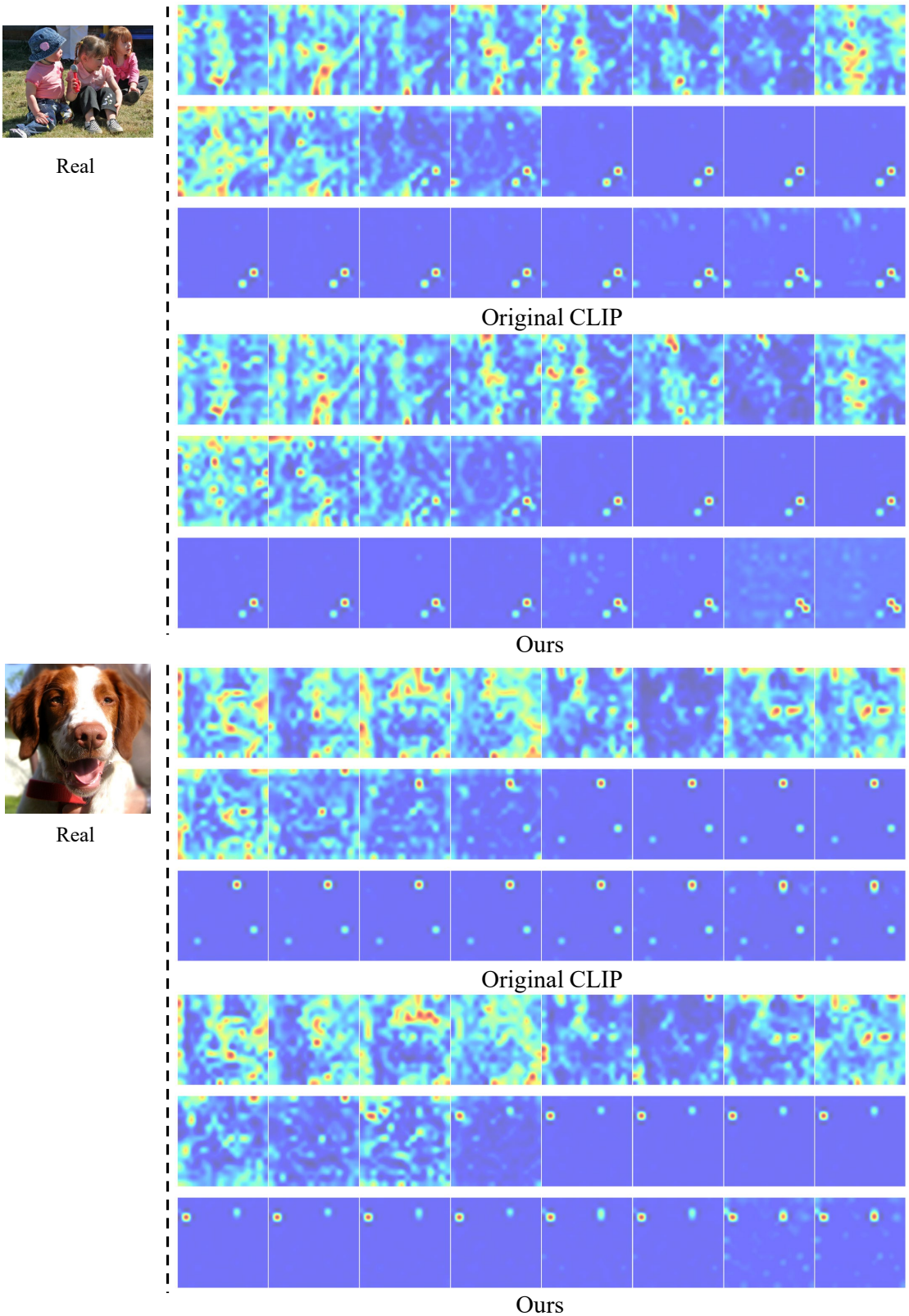

Figure 3. **Self-Attention map** between original CLIP backbone and our finetuned backbone. There are 24 ViT blocks in the image encoder, we plot 8 blocks in each row, with indices increasing from left to right. For clarity of visualization, we use bicubic interpolation between image patches. In the shallow layers, we preserve most semantic features, in deep layers, our attention includes a wider range compared to original CLIP.

References

[1] Midjourney. Inhttps://www.midjourney.com/home/, 2022. 1

[2] Wukong, 2022. 5. Inhttps://xihe.mindspore.cn/modelzoo/wukong, 2022. 5. 1

[3] Adobe. Adobe firefly. https://firefly.adobe.com/, 2025. Accessed: 2025-11-04. 1

[4] Songran Bai, Yuheng Ji, Yue Liu, Xingwei Zhang, Xiaolong Zheng, and Daniel Dajun Zeng. Alleviating performance disparity in adversarial spatiotemporal graph learning under zero-inflated distribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11436–11444, 2025. 4

[5] Shuanghao Bai, Wenxuan Song, Jiayi Chen, Yuheng Ji, Zhide Zhong, Jin Yang, Han Zhao, Wanqi Zhou, Wei Zhao, Zhe Li, et al. Towards a unified understanding of robot manipulation: A comprehensive survey. *arXiv preprint arXiv:2510.10903*, 2025. 4

[6] Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024. 1, 6

[7] Black Forest Labs. FLUX.1: Speeding up text-to-image generation. https://blackforestlabs.ai, 2025. Accessed: 2025-11-26. 4

[8] Andrew Brock et al. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 1

[9] 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. 2, 5, 6

[10] 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. 2, 5, 6

[11] Yunjey Choi et al. 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. 1

[12] Katherine Crowson, Stefan Andreas Baumann, Alex Birch, Tanishq Mathew Abraham, Daniel Z Kaplan, and Enrico Shippole. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In *Proceedings of the 41st International Conference on Machine Learning*, pages 9550–9575. PMLR, 2024. 1

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[14] Prafulla Dhariwal et al. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1

[15] Google, Inc. Nano banana. https://www.nano-banana.com/, 2025. Accessed: 2025-08-29. 4

[16] 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. 1

[17] Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 18685–18694, 2025. 2, 5, 6

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 2

[19] Yuheng Ji, Yue Liu, Zhicheng Zhang, Zhao Zhang, Yuting Zhao, Xiaoshuai Hao, Gang Zhou, Xingwei Zhang, and Xiaolong Zheng. Enhancing adversarial robustness of vision-language models through low-rank adaptation. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, pages 550–559, 2025. 4

[20] Yuheng Ji, Huajie Tan, Cheng Chi, Yijie Xu, Yuting Zhao, Enshen Zhou, Huaihai Lyu, Pengwei Wang, Zhongyuan Wang, Shanghang Zhang, et al. Mathsticks: A benchmark for visual symbolic compositional reasoning with matchstick puzzles. *arXiv preprint arXiv:2510.00483*, 2025. 4

[21] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1724–1734, 2025. 4

[22] Yuheng Ji, Yipu Wang, Yuyang Liu, Xiaoshuai Hao, Yue Liu, Yuting Zhao, Huaihai Lyu, and Xiaolong Zheng. Visualtrans: A benchmark for real-world visual transformation reasoning. *arXiv preprint arXiv:2508.04043*, 2025. 4

[23] Tero Karras et al. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 1

[24] Tero Karras et al. 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. 1

[25] Kvikontent. Kvikontent-midjourney v6. https://huggingface.co/Kvikontent/midjourney-v6, 2023. 1

[26] Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2405–2414, 2025. 2, 5, 6

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. Submitted November 14, 2017; revised January 4, 2019. 2

[28] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang

Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556*, 2023. 1

[29] Huaihai Lyu, Chaofan Chen, Yuheng Ji, and Changsheng Xu. Egoprompt: Prompt pool learning for egocentric action recognition. *arXiv preprint arXiv:2508.03266*, 2025. 4

[30] Midjourney, Inc. Midjourney v6. https://www.midjourney.com, 2025. AI model version 6.0, Accessed: 2025-11-26. 4

[31] 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. 1

[32] 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. 1, 2, 5, 6

[33] 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 (CVPR)*, pages 8245–8257, 2025. 1, 5, 6

[34] Taesung Park et al. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 1

[35] Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023. 1

[36] 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 (CVPR)*, pages 10684–10695, 2022. 1, 2

[37] Andreas Rossler et al. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019. 1

[38] Arseniy Shakhmatov, Anton Razzhigaev, Aleksandr Nikolich, Vladimir Arkhipkin, Igor Pavlov, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky 2.2. https://github.com/ai-forever/Kandinsky-2, 2023. 1

[39] Zirui Song, Guangxian Ouyang, Mingzhe Li, Yuheng Ji, Chenxi Wang, Zixiang Xu, Zeyu Zhang, Xiaoqing Zhang, Qian Jiang, Zhenhao Chen, et al. Maniplvm-r1: Reinforcement learning for reasoning in embodied manipulation with large vision-language models. *arXiv preprint arXiv:2505.16517*, 2025. 4

[40] 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. 1, 2, 5, 6

[41] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 4

[42] Huajie Tan, Cheng Chi, Xiansheng Chen, Yuheng Ji, Zhongxia Zhao, Xiaoshuai Hao, Yaoxu Lyu, Mingyu Cao, Junkai Zhao, Huaihai Lyu, et al. Roboos-next: A unified memory-based framework for lifelong, scalable, and robust multi-robot collaboration. *arXiv preprint arXiv:2510.26536*, 2025. 4

[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. 1

[44] Ming Tao, Bing-Kun Bao, Hao Tang, and Changsheng Xu. Galip: Generative adversarial clips for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14214–14223, 2023. 1

[45] BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, et al. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025. 4

[46] DeciAI Research Team. Decidiffusion 2.0, 2024. 1

[47] 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. 1, 2, 5, 6

[48] Yipu Wang, Yuheng Ji, Yuyang Liu, Enshen Zhou, Ziqiang Yang, Yuxuan Tian, Ziheng Qin, Yue Liu, Huajie Tan, Cheng Chi, Zhiyuan Ma, Daniel Dajun Zeng, and Xiaolong Zheng. Towards cross-view point correspondence in vision-language models. *arXiv preprint arXiv:2512.04686*, 2025. 4

[49] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024. 1, 2, 5, 6

[50] Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. D3: Scaling up deepfake detection by learning from discrepancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 5, 6

[51] Jun-Yan Zhu et al. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 1

[52] 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. 1