CausalCLIP: Causally-Informed Feature Disentanglement and Filtering for Generalizable Detection of Generated Images

Bo Liu[1, 2], **Qiao Qin**[1, 3], **Qinghui He**[1, 3*]

[1] Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing, China
[2]School of Artificial Intelligence
[3]School of Computer Science and Technology
boliu@cqupt.edu.cn, s240201048@stu.cqupt.edu.cn, d250201011@stu.cqupt.edu.cn

Abstract

The rapid advancement of generative models has increased the demand for generated image detectors capable of generalizing across diverse and evolving generation techniques. However, existing methods, including those leveraging pre-trained vision-language models, often produce highly entangled representations, mixing task-relevant forensic cues (causal features) with spurious or irrelevant patterns (non-causal features), thus limiting generalization. To address this issue, we propose CausalCLIP, a framework that explicitly disentangles causal from non-causal features and employs targeted filtering guided by causal inference principles to retain only the most transferable and discriminative forensic cues. By modeling the generation process with a structural causal model and enforcing statistical independence through Gumbel-Softmax-based feature masking and Hilbert-Schmidt Independence Criterion (HSIC) constraints, CausalCLIP isolates stable causal features robust to distribution shifts. When tested on unseen generative models from different series, CausalCLIP demonstrates strong generalization ability, achieving improvements of 6.83% in accuracy and 4.06% in average precision over state-of-the-art methods.

Introduction

The rapid development of generative models such as Generative Adversarial Networks (GANs) (Goodfellow et al. 2014; Karras et al. 2017a) and diffusion models (Ho, Jain, and Abbeel 2020) has drastically lowered the barrier to producing high-quality generated images. While these technologies hold great potential in applications such as image generation, editing, and enhancement, they also pose serious societal risks. Misuse of generative techniques can lead to the creation of hyper-realistic forged content, including manipulated faces, counterfeit evidence, and disinformation, which threatens public security, undermines media credibility, and challenges social governance. This growing threat has sparked an urgent need for a reliable and generalizable detector capable of identifying generated images across a wide range of generative models.

Early AI-generated image detection methods typically relied on supervised training of CNN-based classifiers (Wang et al. 2020; Frank et al. 2020), where models learned generator-specific artifacts (e.g., upsampling traces or frequency anomalies). While these methods performed well on known generators, they exhibit severe performance degradation when facing unseen generation techniques due to overfiting to dataset-specific cues. To address this, recent research has explored leveraging pre-trained models, especially vision-language models like CLIP (Radford et al. 2021), which provide rich and semantically alignment representations with improved cross-domain transferability. Methods like CLIPping (Khan and Dang-Nguyen 2024; Cozzolino et al. 2024) adapt these features through linear probes, prompt tuning, or lightweight adapters to boost detection accuracy. However, even these methods still operate in highly entangled feature spaces, where causal forensic cues are mixed with spurious or non-causal patterns. VIB-Net (Zhang et al. 2025) attempts to alleviate this is-

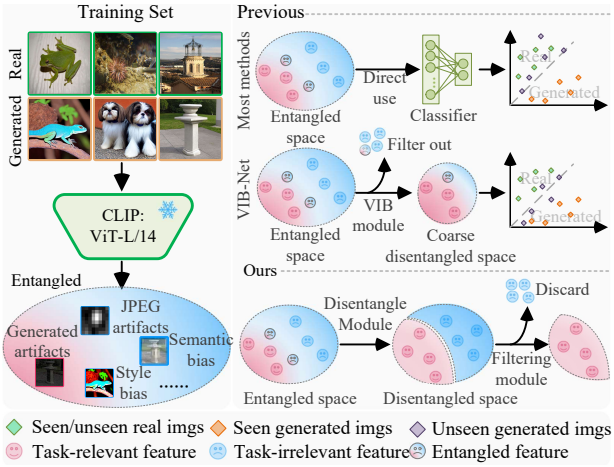

Figure 1: Comparison with previous methods for generated image detection. Prior approaches either directly use entangled CLIP features or filter them in the original space, leading to overfitting to spurious cues. Our method disentangles features into causal and non-causal features and performs filtering in the disentangled space, enabling better preservation of robust forensic cues and improved generalization to unseen generators.

*Corresponding author

sue by applying an information bottleneck to suppress task-irrelevant features, yet it does so without explicitly disentangling causal and non-causal features, leading to coarse filtering and suboptimal generalization under distribution shifts.

In this paper, we find that achieving strong cross-generator generalization requires explicitly separating causal forensic cues from spurious or training data-specific artifacts, rather than merely suppressing irrelevant features in an entangled representation space, which is a coarse filtering strategy that risks discarding task-relevant features, as shown in Fig. 1. Building on this insight, we propose CausalCLIP, a causally guided framework that first disentangles causal features from non-causal features and then selectively leverages the causal subspace for detection. This disentangle-then-filter paradigm preserves stable and transferable forensic evidence that remains effective across diverse types of generative models, providing a theoretically grounded solution to overcoming the limitations of prior methods. Experimental results show that the proposed method achieves strong detection performance with remarkable cross-model generalization, demonstrating its effectiveness across a wide range of generative models.

We summarize our key contributions as follows:

- We propose CausalCLIP, a detection framework following a disentangle-then-filter paradigm, which separates task-relevant features from task-irrelevant features to achieve stronger cross-model generalization.
- The proposed framework leverages adversarial disentanglement and counterfactual interventions to suppress non-causal features and preserve stable forensic cues for robust detection.
- When tested on unseen generative models from different series, CausalCLIP demonstrates strong generalization ability, achieving improvements of 6.83% in accuracy and 4.06% in average precision over state-of-the-art methods.

Related Works

Generated Image Detection

Although GANs (e.g., BigGAN (Brock, Donahue, and Simonyan 2018), ProGAN (Karras et al. 2017b), and StyleGAN2/3 (Karras et al. 2020, 2021)) have achieved remarkable image quality, they still leave subtle forensic cues, such as checkerboard patterns caused by upsampling and color inconsistencies. In contrast, diffusion models (e.g., ADM (Dhariwal and Nichol 2021), Glide (Nichol et al. 2021), and Stable Diffusion (Rombach et al. 2022)) generate highly photorealistic images through iterative denoising, producing much fewer obvious artifacts but exhibiting distinct traces, such as over-smoothed textures and exaggerated style. The diversity and subtlety of these artifacts pose significant challenges for developing detectors that can generalize across different generative models.

Generator-specific Detection Early methods for detecting AI-generated images primarily relied on supervised CNN-based classifiers, which learn to exploit generator-specific artifacts such as upsampling traces (Wang et al. 2020; Durall, Keuper, and Keuper 2020), frequency anomalies (Frank et al. 2020; Chandrasegaran, Tran, and Cheung 2021; Liu et al. 2021, 2022; Bi et al. 2023), and color inconsistencies (Barni et al. 2020; Chandrasegaran et al. 2022). Some more recent methods (Tan et al. 2024; Jia et al. 2025) also follow this strategy, leveraging visual artifacts or generator-specific cues to distinguish real and generated images. While these methods achieve high accuracy on known generators, they often fail to generalize to unseen generative models due to their strong dependence on specific features tied to the training distribution.

Universal Generated Image Detection To further improve cross-generator generalization, recent research has turned to features extracted from large-scale pre-trained models, particularly CLIP (Radford et al. 2021), as universal feature extractors. UnivFD (Ojha, Li, and Lee 2023) demonstrates that CLIP embeddings combined with lightweight classifiers can surpass previous detectors across both GAN-based and diffusion-based datasets. Raising (Cozzolino et al. 2024) further reveals that frozen CLIP features, even when paired with simple classifiers, exhibit strong zero- and few-shot generalization. Building on this paradigm, CLIPping (Khan and Dang-Nguyen 2024) adapts CLIP via prompt tuning and adapter-based fine-tuning, while C2P-CLIP (Tan et al. 2025) injects category-level prompts to better align image-text representations, thereby improving cross-domain detection performance. Moreover, the Few-Shot Learner (Wu et al. 2025) extends this line of research by learning specialized metric-based detectors using only a handful of samples from a new generator. Despite these advances, CLIP-based methods still operate on entangled feature space, where task-relevant features are mixed with task-irrelevant features. VIB-Net (Zhang et al. 2025) attempts to alleviate this by introducing an information bottleneck to suppress irrelevant features, but its filtering remains coarse and lacks explicit feature disentanglement, resulting in limited cross-generator generalization.

Representation Learning

Causal representation learning (CRL) aims to improve generalization under distribution shifts by learning representations that align with underlying causal factors rather than spurious correlations (Schölkopf et al. 2021). In various domains, CRL has proven effective in disentangling task-relevant signals from irrelevant noise, thereby enhancing model interpretability and transferability. Typical approaches include adversarial training for disentanglement (Mathieu et al. 2016), structural causal models (SCMs) for explicitly modeling latent causal structures (Yang et al. 2021), and counterfactual interventions to reduce dataset bias and isolate stable features. For instance, CausalVAE (Yang et al. 2021) leverages causal mechanisms to construct endogenous variables from exogenous noise, while DEAR (Shen et al. 2022) combines GANs with SCMs under weak supervision to relax independence assumptions and model causally entangled factors. These findings suggest that disentangling causal and non-causal features could help detection models focus on stable forensic evidence that re-

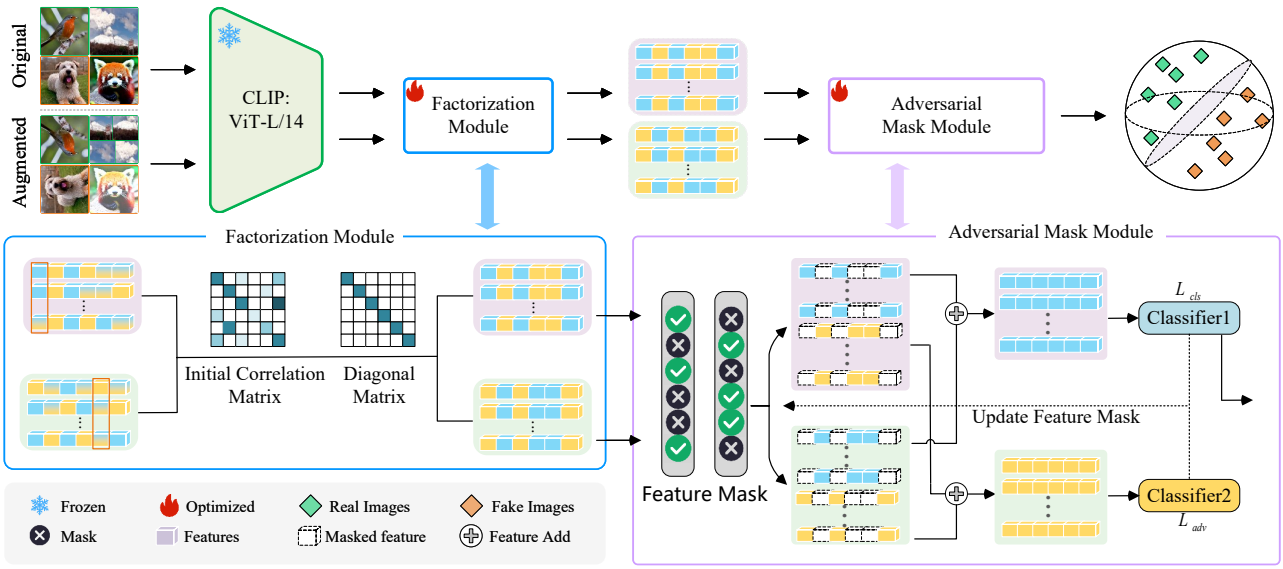

Figure 2: Architecture of the proposed CausalCLIP framework. An input image is processed by CLIP-ViT-L/14 to extract frozen features, which are disentangled by a Factorization Module into causal and non-causal components via correlation analysis. A Feature Mask is then learned by the Adversarial Mask Module to suppress non-causal features. Two binary classifiers are used: Classifier 1 predicts real vs. fake from masked features, while Classifier 2 acts adversarially, attempting to classify using the masked-out features. The mask is optimized to aid Classifier 1 while minimizing Classifier 2's success, promoting the retention of stable, generation-invariant forensic features for better generalization.

mains effective across different generative models, addressing the common issue that features predictive on one generator often become unreliable when the generative mechanism changes.

Method

In this section, we introduce CausalCLIP, a causally-informed framework for generated image detection. We first provide an overview of the entire pipeline, as illustrated in Figure 2, and then describe each core module in detail. Specifically, we present the feature disentanglement (factorization) module, which separates causal and non-causal components of the extracted features, followed by the adversarial masking module, which enforces causal invariance through counterfactual interventions. Finally, we formulate the overall training objective and optimization strategy.

Overview

The key idea of CausalCLIP is to improve cross-model generalization in generated image detection by explicitly separating stable, task-relevant forensic cues (causal features) from non-causal features that are tied to specific datasets or generation styles. As illustrated in Figure 2, the input image is first processed by a frozen CLIP encoder to extract high-level semantic features. These features are then fed into the factorization module, which separates them into two complementary parts: causal features that capture intrinsic forensic cues stable across generation mechanisms, and non-causal features that encode generator- or dataset-specific artifacts. To further enhance generalization, the adversarial masking module performs targeted interventions on the non-causal features, weakening their influence and encouraging the classifier to focus on the causal features that remain reliable under distribution shifts. Finally, the refined causal features are passed to a lightweight classifier to determine whether the input image is real or generated.

Factorization Module

A major challenge in generated image detection is that CLIP embeddings entangle causal features, those truly indicative of real vs. fake, with non-causal artifacts specific to generators or datasets. This entanglement leads to overfitting and poor generalization to unseen generators. We address this with a Factorization Module that explicitly separates stable causal features from variable non-causal ones.

We assume that an image $X \in \mathcal{X}$ can be explained by a structural causal model with two independent factors:

$$Z_c := f_c(G, \epsilon_c), \tag{1}$$

$$Z_{nc} := f_{nc}(C, \epsilon_{nc}), \tag{2}$$

$$X := g(Z_c, Z_{nc}, \epsilon_x), \tag{3}$$

where Z_c represents causal features, while Z_{nc} captures non-causal features. Here, G and C are latent variables related to generation-independent content factors and generator-specific style or artifact factors, respectively, and ϵ. denotes exogenous noise. Our objective is to recover Z_c from the entangled CLIP features while suppressing the influence of Z_{nc}.

Given the CLIP embedding $E = \text{CLIP}(X) \in \mathbb{R}^d$, the factorization module learns a feature mask $M \in [0, 1]^d$ to separate:

$$\tilde{Z}_c = M \odot E, \quad \tilde{Z}_{nc} = (1 - M) \odot E, \tag{4}$$

where $\odot$ denotes element-wise multiplication. The mask M is parameterized by a Gumbel-softmax function:

$$M = \sigma((\text{MLP}(E) + g)/\tau), \quad g \sim \text{Gumbel}(0, 1), \tag{5}$$

with temperature τ controlling the sparsity of feature selection. This mechanism ensures differentiable feature selection and yields a clean causal subspace $\tilde{Z}_c$ for downstream classification.

Adversarial Masking Module

Although the factorization module separates CLIP embeddings into causal and non-causal features, residual non-causal signals may still affect the classifier. To address this, we introduce an Adversarial Masking Module that uses an adversarial mechanism to ensure the decision boundary relies only on the stable causal subspace.

The adversarial design is motivated by the observation that explicitly suppressing non-causal features improves generalization under distribution shifts. We set up a minimax game where:

- A classifier h predicts $Y \in \{0, 1\}$ (real vs. generated) based on the causal features $\tilde{Z}_c$.
- An adversary d attempts to predict Y from the non-causal features $\tilde{Z}_{nc}$.

The classifier and mask M are optimized to minimize classification loss while making $\tilde{Z}_{nc}$ uninformative, forcing the model to rely solely on $\tilde{Z}_c$.

The classification loss is defined as the standard binary cross-entropy:

$$\mathcal{L}_{cls} = -\mathbb{E}_{X,Y}[Y \log h(\tilde{Z}_c) + (1-Y) \log(1-h(\tilde{Z}_c))]. \tag{6}$$

where $h(\tilde{Z}_c)$ denotes the predicted probability that X is generated.

The adversary is trained to maximize its ability to recover Y from non-causal features, while the mask and classifier are optimized to suppress such information. The adversarial loss is:

$$\mathcal{L}^{+}_{adv} = -\mathbb{E}_{X,Y}[Y \log d(\tilde{Z}_{nc})], \tag{7}$$

$$\mathcal{L}^{-}_{adv} = -\mathbb{E}_{X,Y}[(1-Y) \log(1 - d(\tilde{Z}_{nc}))]. \tag{8}$$

$$\mathcal{L}_{adv} = \mathcal{L}^{+}_{adv} + \mathcal{L}^{-}_{adv}. \tag{9}$$

To encourage a clear separation, we regularize the mask with sparsity and independence terms:

$$\mathcal{L}_{\text{mask}} = \lambda_1 \|M\|_1 + \lambda_2 \widehat{\text{HSIC}}(\tilde{Z}_c, \tilde{Z}_{nc}), \tag{10}$$

where the ℓ_1-norm promotes sparse feature selection, and the empirical HSIC term encourages statistical independence between the causal and non-causal subspaces. The HSIC is computed using Gaussian kernels with bandwidth selected via the median heuristic.

To enhance robustness, we introduce counterfactual interventions on causal features. By randomly masking a subset of dimensions,

$$\tilde{Z}^{CF}_c = \tilde{Z}_c \odot (1 - B), \quad \text{where } B \sim \text{Bernoulli}(p), \tag{11}$$

we simulate distributional perturbations and enforce prediction consistency via

$$\mathcal{L}_{\text{inv}} = \text{KL}\big(h(\tilde{Z}_c) \parallel h(\tilde{Z}^{CF}_c)\big), \tag{12}$$

compelling the classifier to rely on stable causal semantics rather than generator-dependent cues.

Optimization Objective

Our goal is to learn representations that preserve stable causal features while suppressing non-causal signals across generative models. We unify classification, adversarial, mask regularization, and counterfactual intervention losses as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} - \alpha \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{mask}} + \beta \mathcal{L}_{\text{inv}}, \tag{13}$$

where α and β balance adversarial disentanglement and counterfactual consistency.

The objective is guided by two principles: (1) **Disentanglement**, where $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{mask}}$ separate causal from non-causal features; (2) **Stability**, where $\mathcal{L}_{\text{inv}}$ enforces consistent predictions under causal perturbations. Training alternates: h minimizes $\mathcal{L}_{\text{cls}} + \beta \mathcal{L}_{\text{inv}}$, d maximizes $\mathcal{L}_{\text{adv}}$, and M minimizes $\mathcal{L}_{\text{total}}$, progressively refining causal features and suppressing non-causal ones.

Experiments

In this section, we evaluate the proposed CausalCLIP framework on various datasets and compare it with SOTA generated image detection methods. We first introduce the training and testing datasets and evaluation metrics, followed by implementation details.

Experimental Setting

Training Datasets To evaluate the detection performance and generalization performance of the proposed method across different generative models. We use two representative datasets for training. The first is ProGAN from CNNDet (Wang et al. 2020), which contains 360k real images from LSUN (Yu et al. 2015) and 360k generated images from ProGAN. The second is Stable Diffusion v1.4 from GenImage (Zhu et al. 2023), which includes 32k high-quality generated images from Stable Diffusion v1.4 (Rombach et al. 2022), paired with their corresponding real images. All methods are trained on either ProGAN or SDv1.4 datasets to evaluate cross-generator generalization.

Testing Datasets We consider 15 testing datasets covering both GAN-based and diffusion-based models. The ForenSynths test set includes images generated by six different GAN-based models, representing a variety of architectures such as ProGAN (Goodfellow et al. 2014), CycleGAN (Zhu et al. 2017), StarGAN (Choi et al. 2018), StyleGAN (Zhu et al. 2017), BigGAN (Brock, Donahue, and Simonyan 2018), and GauGAN (Li et al. 2020). In addition, two other models, DeepFake (Rossler et al. 2019a) and SAN (Dai et al. 2019), are also included. The corresponding real images are sampled from six commonly used datasets, namely ImageNet (Russakovsky et al. 2015), LSUN (Yu et al. 2015),

Method	Diffusion Models							Generative Adversarial Networks						Others		*AP*
	SD1.4	SD1.5	ADM	GLIDE	Midj	Wukong	VQDM	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Star-GAN	Gau-GAN	Deep-fake	SAN	Avg
CNNSpot	99.98	99.83	51.10	58.80	67.93	99.80	49.92	53.15	50.23	50.22	49.79	47.07	56.08	54.86	54.03	63.24
Fusing	99.90	97.98	69.30	94.20	81.20	99.90	84.60	67.63	87.79	64.84	69.37	91.20	43.08	72.56	89.42	81.07
Lgrad	99.94	99.92	58.52	84.00	91.06	99.72	56.34	83.59	90.24	56.49	47.51	69.93	49.25	66.49	65.09	78.24
Univfd	96.04	96.26	66.34	93.73	91.06	90.98	74.53	51.77	63.42	72.00	75.81	54.12	65.99	70.24	83.34	75.31
NPR	**100.00**	99.97	94.70	95.80	95.50	**100.00**	86.30	83.30	94.90	89.60	72.00	82.70	66.00	**85.30**	95.90	89.98
CLIPping	93.97	93.10	68.00	87.44	77.34	86.52	77.17	88.54	88.44	61.64	85.33	77.23	81.56	61.19	57.37	80.87
VIB-Net	**100.00**	99.97	**95.49**	97.13	97.81	99.93	83.02	96.59	**98.44**	81.50	**97.17**	84.31	**96.94**	81.32	93.27	94.60
Ours	**100.00**	**99.99**	95.37	**99.47**	**98.23**	99.95	**98.82**	**97.33**	98.35	**97.43**	96.29	**98.62**	96.84	79.35	**97.79**	**96.92**

Table 1: The *AP* values of different methods trained on Diffusion source images. Data in bold represents the best, while data underlined represents the second best.

Method	Diffusion Models							Generative Adversarial Networks						Others		*ACC*
	SD1.4	SD1.5	ADM	GLIDE	Midj	Wukong	VQDM	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Star-GAN	Gau-GAN	Deep-fake	SAN	Avg
CNNSpot	99.48	93.35	50.10	50.90	56.42	97.90	50.04	50.27	49.81	50.10	50.98	49.77	50.38	51.98	50.22	60.51
Fusing	99.90	**99.91**	51.30	57.50	52.30	99.90	64.20	51.20	52.40	53.50	50.20	58.20	49.32	51.02	64.48	63.71
Lgrad	99.12	99.05	53.00	64.24	76.34	97.53	50.93	61.61	60.74	48.82	61.43	50.17	49.70	50.17	56.49	65.29
Univfd	83.55	84.80	53.35	75.30	71.60	73.55	55.10	58.65	59.30	61.45	56.80	61.45	55.30	58.40	72.00	65.37
NPR	**100.00**	99.90	73.00	**89.70**	82.30	**100.00**	68.30	60.30	67.20	59.20	58.00	73.20	52.00	**74.80**	**89.60**	76.50
CLIPping	96.07	95.48	70.14	85.00	77.66	88.86	79.35	88.78	88.48	89.57	80.69	90.82	85.94	66.91	61.64	83.03
VIB-Net	99.55	99.20	73.85	74.25	**88.05**	98.25	89.35	89.70	88.60	91.20	74.10	80.70	87.15	72.00	81.50	85.83
Ours	99.81	99.59	**77.95**	89.22	80.72	99.08	**90.73**	**91.22**	**97.35**	**94.27**	**87.78**	**95.75**	**94.36**	71.64	87.21	**90.45**

Table 2: The *ACC* values of different methods trained on Diffusion source images. Similar to the symbols in Table 1.

FaceForensics++ (Rossler et al. 2019b), CelebA-HQ (Karras et al. 2017a), CelebA (Liu et al. 2015), and COCO (Lin et al. 2014). The GenImage test set contains images generated by seven state-of-the-art diffusion models, including Stable Diffusion v1.4, v1.5 (Rombach et al. 2022), VQDM (Gu et al. 2022), ADM (Dhariwal and Nichol 2021), GLIDE (Nichol et al. 2021), Wukong, and Midjourney. Each generated image is paired with a corresponding real image to ensure fair and comparable evaluation.

Evaluation Metrics We adopt two commonly used metrics: Average Precision (*AP*) and Accuracy (*ACC*). *AP* is a threshold-independent metric that provides a comprehensive evaluation of the model's performance across different decision thresholds. In contrast, *ACC* measures the proportion of correctly classified samples among all test samples and is sensitive to the choice of classification threshold.

Baselines We compare the proposed CausalCLIP with a diverse set of SOTA detection methods, including CNNDetection (CVPR 2020)(Wang et al. 2020), FreDect (ICML 2020)(Frank et al. 2020), Fusing (ICIP 2022)(Ju et al. 2022), LGrad (CVPR 2023)(Tan et al. 2023), DIRE (CVPR 2023)(Wang et al. 2023), UnivFD (CVPR 2023)(Ojha, Li, and Lee 2023), NPR (CVPR 2024)(Tan et al. 2024), CLIPping (ICMR 2024)(Khan and Dang-Nguyen 2024), and VIB-Net (CVPR 2025) (Zhang et al. 2025).

Implementation Details We employ the frozen image encoder of the pretrained CLIP-ViT/L-14 model as the backbone to extract high-level semantic features. On top of the extracted features, we train the downstream causal representation module and the discriminative classifier. We use the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 256, with early stopping applied to prevent overfitting. All input images are center-cropped to 224×224 pixels. The entire framework is implemented in PyTorch, and all experiments are conducted on a NVIDIA Tesla V100 GPU.

Cross-Model Generalization Evaluation

To comprehensively assess the generalization capability of CausalCLIP under distribution shifts, we design two challenging cross-model detection experiments. Specifically, we train the model on images generated by either a diffusion model or a GAN, and test it on a diverse set of generative models. This setup mimics real-world scenarios where training and testing distributions differ significantly due to variations in generation mechanisms and visual styles.

Diffusion-Sources Evaluation In the first setting, we train CausalCLIP on images generated by Stable Diffusion v1.4

Methods	Generative Adversarial Networks						Others		Diffusion Models							*AP*
	Pro-GAN	Cycle-GAN	Big-GAN	Stytle-GAN	Star-GAN	Gau-GAN	Deep-fake	SAN	SD1.4	SD1.5	ADM	GLIDE	Midj	Wukong	VQDM	Avg
CNNSpot	99.99	96.40	87.50	96.94	94.24	98.28	64.42	55.89	52.86	53.25	65.14	68.10	56.60	51.15	69.49	74.02
FreDect	99.99	84.77	93.62	88.97	99.48	82.85	70.77	49.50	38.50	38.41	63.72	54.73	47.25	40.44	86.01	69.27
Fusing	**100.00**	95.50	90.76	99.48	99.82	88.32	71.12	77.33	65.30	65.62	74.85	77.44	69.91	64.53	75.42	81.03
Lgrad	99.90	94.01	90.75	**99.80**	**99.98**	79.29	71.71	45.09	70.90	71.72	71.83	75.96	71.42	66.51	70.23	78.61
Univfd	**100.00**	99.21	98.31	97.98	99.35	99.80	82.04	82.18	85.48	82.30	84.34	84.04	69.10	90.13	94.96	89.85
NPR	**100.00**	98.50	87.80	99.80	99.90	85.50	82.40	71.60	84.00	84.60	74.60	85.70	**85.40**	80.50	81.20	86.77
CLIPping	99.85	94.03	91.27	95.24	99.05	89.52	73.89	56.62	58.28	57.85	77.25	77.93	52.43	60.65	80.00	77.59
VIB-Net	**100.00**	**99.80**	99.29	98.79	99.72	**99.99**	**92.64**	91.62	87.24	**86.98**	87.88	88.53	75.68	90.92	96.51	93.04
Ours	**100.00**	99.27	**99.31**	99.70	99.89	99.72	90.78	**93.14**	**89.91**	87.59	**88.92**	**92.41**	82.72	**93.52**	**97.12**	**94.27**

Table 3: The *AP* values of different methods trained on GAN source images. Similar to the symbols in Table 1.

Methods	Generative Adversarial Networks						Others		Diffusion Models							*ACC*
	Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Star-GAN	Gau-GAN	Deep-fake	SAN	SD1.4	SD1.5	ADM	GLIDE	Midj	Wukong	VQDM	Avg
CNNSpot	99.99	87.59	71.18	89.95	94.60	81.44	51.69	50.00	50.82	50.88	60.20	57.85	50.77	51.13	56.20	66.95
FreDect	99.36	78.76	81.97	78.01	94.62	80.57	63.29	50.00	40.02	40.38	64.66	55.43	46.89	41.54	78.95	66.30
Fusing	99.90	87.01	77.32	85.21	97.04	76.95	53.76	54.56	51.04	51.34	56.50	57.15	52.12	51.67	55.09	67.11
Lgrad	99.80	86.94	85.63	91.08	99.27	72.49	56.42	44.47	63.03	63.67	67.10	66.10	56.20	63.60	67.02	72.19
Univfd	99.90	98.50	94.50	84.40	95.85	99.50	67.40	56.50	63.10	63.57	66.90	61.70	57.85	71.06	85.00	77.72
NPR	99.80	96.10	84.40	**97.70**	99.30	82.50	80.20	69.20	76.60	77.90	69.70	77.30	**77.80**	76.10	78.10	82.85
CLIPping	99.88	96.74	94.77	94.87	99.47	94.15	76.48	60.27	61.26	60.78	**80.77**	80.69	53.95	63.94	84.76	80.19
VIB-Net	99.99	99.00	**95.75**	91.25	98.95	**99.70**	**83.20**	70.50	71.55	70.00	71.45	69.40	61.25	75.90	86.65	82.97
Ours	99.06	**99.95**	91.73	92.73	96.14	95.92	72.34	**79.45**	**82.58**	**82.54**	78.94	**82.73**	65.99	**85.91**	**87.48**	**86.23**

Table 4: The *ACC* values of different methods trained on GAN source images. Similar to the symbols in Table 1.

and evaluate its performance on samples from 15 other generative models. As shown in Tables 1 and 2, many existing detection methods (e.g., LGrad, UnivFD) suffer substantial performance drops on these newer diffusion models, with declines exceeding 40% in both *AP* and *ACC*. This indicates that these methods often overfit to non-causal artifacts in the training data, such as stylistic inconsistencies or generator-specific textures, which fail to generalize under distributional shifts.

In contrast, CausalCLIP consistently maintains strong performance across all models, achieving improvements of 2.32% in *AP* and 4.62% in *ACC*. When further tested on unseen GAN-based generators, it achieves additional improvements of 6.83% in *ACC* and 4.06% in *AP*, confirming strong cross-family generalization. These gains are attributed to CausalCLIP's causality-guided representation learning, where the causal mask and adversarial intervention collaboratively disentangle causal features and foster counterfactually invariant representations for robust generalization.

GAN-Sources Evaluation To further assess the generalization ability of CausalCLIP under pronounced style variations, we conduct a second experiment where the model is trained on images generated by ProGAN and evaluated on 15 diverse generative models. Due to the substantial differences in generative characteristics between ProGAN and modern diffusion models, this setting introduces a more severe distribution gap, making it well-suited for examining the model's robustness to both style and semantic shifts.

Results in Tables 3 and 4 demonstrate that many baseline methods (e.g., Fusing, LGrad, UnivFD) suffer severe performance drops on the new diffusion models, with *AP* and *ACC* often falling below 60% and, in some cases, plummeting to 55% on models such as ADM, Midjourney, and VQDM. These results highlight the heavy reliance of conventional methods on generator-specific artifacts, which fail to transfer across different generation paradigms. In contrast, CausalCLIP consistently outperforms all baselines, achieving improvements of 1.23% in *AP* and 3.26% in *ACC*. When further tested on unseen Diffusion-based generators, it achieves additional improvements of 8.57% in *ACC* and 2.64% in *AP*, confirming strong generalization.

Feature Disentanglement

Feature Space Visualization To assess the discriminative power of different representations, we visualize fea-

Factorization module	Masking module	*ACC*		*AP*	
		OP	GP	OP	GP
×	×	65.37	60.42	75.31	67.09
✓	×	79.42	75.91	89.53	87.78
×	✓	70.73	65.94	82.13	79.28
✓	✓	90.45	89.95	96.92	95.25

Table 5: Ablation study on the contributions of the disentanglement and masking modules. "×" indicates the module is disabled, while "✓" indicates it is enabled. The baseline UnivFD corresponds to the case where both modules are disabled. Combining both modules achieves the best results.

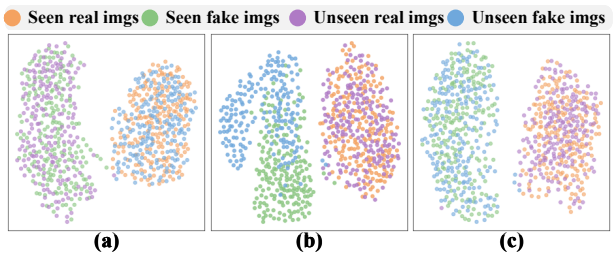

Figure 3: UMAP visualization of real and fake image features under seen and unseen settings. (a) CLIP shows strong domain entanglement, (b) VIB achieves partial separation, (c) our method provides clear separation across all domains.

tures using UMAP for both seen and unseen generators. As shown in Figure 3, CLIP features exhibit strong entanglement across domains, making real–fake discrimination difficult. VIB achieves partial separation but still overlaps in unseen cases. In contrast, our method achieves clear separation across all domains, indicating superior disentanglement and generalization.

Robustness We further evaluate model robustness against two common input perturbations: JPEG compression and Gaussian blur. Figure 4 reports accuracy and average precision under varying quality factors and blur levels. Conventional approaches degrade significantly as perturbations increase, while our method maintains the most stable performance across all settings.

Ablation Studies

To evaluate the individual and combined contributions of the disentanglement (factorization) and masking modules, we perform ablation experiments on the SDv1.4 training dataset. In this evaluation, overall performance (OP) refers to the average performance across all testing datasets, while generalization performance (GP) denotes the average performance on GAN-based datasets. The results for both *ACC* and *AP* are reported in Table 5.

When both modules are disabled, the framework reduces to the baseline UnivFD, which achieves 65.37% (*ACC*) and 75.31% (*AP*) in overall performance, and 60.42% (*ACC*) and

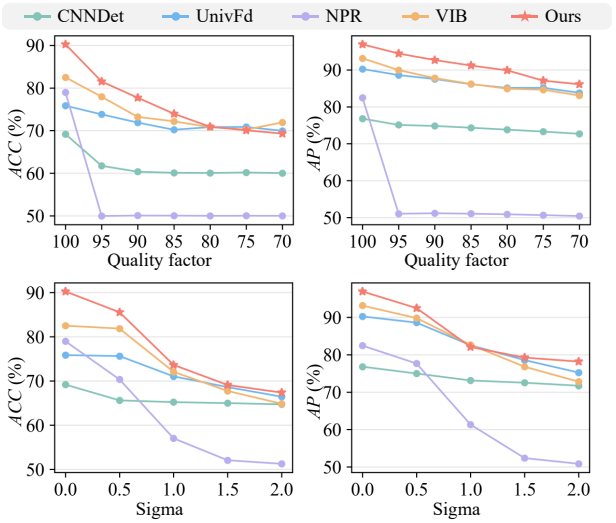

Figure 4: Robustness analysis under JPEG compression (top) and Gaussian blur (bottom). Our method (star markers) shows better stability under most perturbations.

67.09% (*AP*) in generalization performance. Introducing the disentanglement module alone leads to significant improvements of +14.05% (*ACC*) and +14.22% (*AP*) in overall performance compared to the baseline, showing that separating causal and non-causal components helps the model capture task-relevant forensic cues. The masking module alone also brings improvements of +5.36% (*ACC*) and +6.82% (*AP*) in overall performance, demonstrating its ability to suppress spurious correlations. When the two modules are combined, the framework achieves the best results, with 89.64% (*ACC*) and 96.92% (*AP*) in overall performance, and 88.45% (*ACC*) and 95.25% (*AP*) in generalization performance. This corresponds to absolute gains of +24.27% (*ACC*) and +21.61% (*AP*) over the UnivFD baseline. These results confirm that disentanglement and masking are complementary, with disentanglement isolating stable causal features and masking further refining the feature space by removing style-specific noise.

Conclusion

In this paper, we introduce CausalCLIP, a causal representation learning framework aimed at improving the generalization of generated image detectors across diverse generative models. CausalCLIP disentangles representations through a Gumbel-Softmax-based masking mechanism guided by HSIC constraints to isolate causal forensic features from spurious ones. The adversarial masking strategy enables targeted suppression of non-causal components while preserving generalizable cues. Extensive experiments show our approach consistently outperforms existing methods, especially in cross-model settings. These findings highlight the importance of causal feature separation for generalization under distribution shifts. CausalCLIP provides a strong foundation for future image forensics research.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62376046, Grant U24B20182, and Grant 62561160098, and in part by the Natural Science Foundation of Chongqing under Grant CSTB2023NSCQ-MSX0341.

References

Barni, M.; Kallas, K.; Nowroozi, E.; and Tondi, B. 2020. CNN detection of GAN-generated face images based on cross-band co-occurrences analysis. In *2020 IEEE international workshop on information forensics and security (WIFS)*, 1–6. IEEE.

Bi, X.; Liu, B.; Yang, F.; Xiao, B.; Li, W.; Huang, G.; and Cosman, P. C. 2023. Detecting generated images by real images only. *arXiv preprint arXiv:2311.00962*.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Chandrasegaran, K.; Tran, N.-T.; Binder, A.; and Cheung, N.-M. 2022. Discovering transferable forensic features for cnn-generated images detection. In *European Conference on Computer Vision*, 671–689. Springer.

Chandrasegaran, K.; Tran, N.-T.; and Cheung, N.-M. 2021. A closer look at fourier spectrum discrepancies for cnn-generated images detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7200–7209.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Cozzolino, D.; Poggi, G.; Corvi, R.; Nießner, M.; and Verdoliva, L. 2024. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4356–4366.

Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; and Zhang, L. 2019. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11065–11074.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34: 8780–8794.

Durall, R.; Keuper, M.; and Keuper, J. 2020. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7890–7899.

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, 3247–3258. PMLR.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10696–10706.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Jia, Z.; Huang, C.; Zhu, Y.; Fei, H.; Duan, X.; Yuan, Z.; Deng, Y.; Zhang, J.; Zhang, J.; and Zhou, J. 2025. Secret Lies in Color: Enhancing AI-Generated Images Detection with Color Distribution Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13445–13454.

Ju, Y.; Jia, S.; Ke, L.; Xue, H.; Nagano, K.; and Lyu, S. 2022. Fusing global and local features for generalized ai-synthesized image detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, 3465–3469. IEEE.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017a. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017b. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2021. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34: 852–863.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8110–8119.

Khan, S. A.; and Dang-Nguyen, D.-T. 2024. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 1006–1015.

Li, M.; Lin, J.; Ding, Y.; Liu, Z.; Zhu, J.-Y.; and Han, S. 2020. Gan compression: Efficient architectures for interactive conditional gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5284–5294.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, B.; Yang, F.; Bi, X.; Xiao, B.; Li, W.; and Gao, X. 2022. Detecting generated images by real images. In *European Conference on Computer Vision*, 95–110. Springer.

Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; and Yu, N. 2021. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 772–781.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Mathieu, M. F.; Zhao, J. J.; Zhao, J.; Ramesh, A.; Sprechmann, P.; and LeCun, Y. 2016. Disentangling factors of variation in deep representation using adversarial training. *Advances in neural information processing systems*, 29.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24480–24489.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019a. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019b. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1–11.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5): 612–634.

Shen, X.; Liu, F.; Dong, H.; Lian, Q.; Chen, Z.; and Zhang, T. 2022. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241): 1–55.

Tan, C.; Tao, R.; Liu, H.; Gu, G.; Wu, B.; Zhao, Y.; and Wei, Y. 2025. 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, 7184–7192.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024. 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*, 28130–28139.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; and Wei, Y. 2023. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12105–12114.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8695–8704.

Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Hu, H.; Chen, H.; and Li, H. 2023. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22445–22455.

Wu, S.; Liu, J.; Li, J.; and Wang, Y. 2025. Few-Shot Learner Generalizes Across AI-Generated Image Detection. *arXiv preprint arXiv:2501.08763*.

Yang, M.; Liu, F.; Chen, Z.; Shen, X.; Hao, J.; and Wang, J. 2021. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9593–9602.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Zhang, H.; He, Q.; Bi, X.; Li, W.; Liu, B.; and Xiao, B. 2025. Towards Universal AI-Generated Image Detection by Variational Information Bottleneck Network. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23828–23837.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhu, M.; Chen, H.; Yan, Q.; Huang, X.; Lin, G.; Li, W.; Tu, Z.; Hu, H.; Hu, J.; and Wang, Y. 2023. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36: 77771–77782.