Fooling the Watchers: Breaking AIGC Detectors via Semantic Prompt Attacks

Run Hao [1] **Peng Ying** [2,3]

Abstract

The rise of text-to-image (T2I) models has enabled the synthesis of photorealistic human portraits, raising serious concerns about identity misuse and the robustness of AIGC detectors. In this work, we propose an automated adversarial prompt generation framework that leverages a grammar tree structure and a variant of the Monte Carlo tree search algorithm to systematically explore the semantic prompt space. Our method generates diverse, controllable prompts that consistently evade both open-source and commercial AIGC detectors. Extensive experiments across multiple T2I models validate its effectiveness, and the approach ranked first in a real-world adversarial AIGC detection competition. Beyond attack scenarios, our method can also be used to construct high-quality adversarial datasets, providing valuable resources for training and evaluating more robust AIGC detection and defense systems.

1. Introduction

In recent years, text-to-image (T2I) models have witnessed significant advancements, driven by breakthroughs in deep learning and generative modeling. Early works such as AttnGAN (Xu et al., 2018) and StackGAN (Zhang et al., 2017) laid the foundation for generating coherent images from textual descriptions by integrating attention mechanisms and multi-stage refinement strategies. With the emergence of large-scale diffusion models, notably DALL·E 2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), and Stable Diffusion (Rombach et al., 2022), the quality and realism of generated images have drastically improved, enabling high-fidelity synthesis conditioned on natural language prompts.

Alongside the rapid development of T2I models, the prolif-

eration of AI-generated content (AIGC) has raised concerns regarding authenticity and misuse. In particular, their ability to synthesize photorealistic human portraits has introduced new challenges in identity and privacy protection. Hyper-realistic faces can be fabricated with arbitrary attributes, potentially enabling malicious actors to create fake IDs, impersonate real individuals, or even generate sensitive visuals of non-consenting subjects (Karras et al., 2019; Chesney & Citron, 2022; Douglas & Yampolskiy, 2021). Such misuse raises pressing ethical and legal questions, particularly when these synthetic faces are indistinguishable from real ones by both humans and machines (Mirsky & Lee, 2021; Verdoliva, 2020). This has led to growing interest in AIGC detection, which aims to distinguish between real and synthetic content. Recent works in this domain, such as GAN-generated image detectors (Wang et al., 2020a) and diffusion-specific classifiers (Wang et al., 2023), have leveraged both low-level artifacts (e.g., color statistics (Durall et al., 2020), sensor noise (Frank et al., 2020), frequency domain inconsistencies (Durall et al., 2020)) and high-level semantic cues (e.g., unnatural textures (Zhang et al., 2019), unrealistic object layouts, or prompt-image mismatches (Chen et al., 2023; Liang et al., 2023)) to achieve robust detection.

Despite the increasing sophistication of these detectors, many rely on assumptions about known generation models or artifacts, leaving them vulnerable to adaptive or black-box attacks. In high-stakes scenarios involving human portraits, failure to detect synthetic content may lead to severe consequences, including misinformation, reputational harm, or privacy breaches, motivating the need for rigorous evaluation of detector robustness under adversarial conditions. However, conducting such evaluations poses several unique challenges in the context of T2I generation. *First, the inherently stochastic nature of modern T2I models limits the precision with which image content can be controlled.* Unlike traditional adversarial attacks that directly manipulate pixel-level features or inject specific noise patterns, prompt-based attacks must rely on indirect, semantic-level perturbations—often resulting in only approximate control over the generated outputs (Nichol et al., 2021). *Second, the process of manually crafting effective prompts is prohibitively costly and inefficient.* Identifying combinations of textual cues that consistently induce detection failures requires substantial human effort and insight, and it remains difficult to deter-

*Equal contribution [1]Aarhus University [2]School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, China [3]Mine Digitization Engineering Research Center of the Ministry of Education, China University of Mining and Technology, Xuzhou, China. Correspondence to: Run Hao <au785440@uni.au.dk>, Peng Ying <pengying@cumt.edu.cn>.

mine which elements of a prompt are most responsible for the success or failure of an attack.

To overcome these challenges, we develop a grammar tree-based prompt generation mechanism, where each node in the tree corresponds to a semantic component (e.g., portrait attributes or stylistic features), and prompts are constructed through a recursive top-down traversal of this tree. To further optimize this process, we integrate a variant of the Monte Carlo Tree Search (MCTS) algorithm—specifically, the UCT-Rand algorithm (Zheng et al., 2024)—which replaces greedy selection with weighted random sampling. This allows for broader and more effective exploration of the prompt space, increasing the likelihood of generating photorealistic portraits capable of evading both open and closed-source AIGC detectors. To demonstrate the practical effectiveness of our approach, we deployed the proposed method in a real-world AIGC adversarial detection competition (Tencent, 2025), where it achieved first place. Furthermore, we applied our method across multiple state-of-the-art T2I models and evaluated its performance against both open-source and commercial AIGC detectors. Experimental results show that existing detectors are highly susceptible to prompts generated by our framework, further highlighting the fragility of current detection mechanisms when facing prompt-level semantic attacks.

Our contributions are summarized as follows:

- We propose semantic prompt attack methods targeting both closed-source and open-source AIGC detectors, revealing detector's susceptibility to such attacks. To the best of our knowledge, this is the first study on prompt-level semantic attacks against AIGC detectors.

- We design a hierarchical grammar tree to structure semantic segments of prompts, such as portrait attributes and stylistic features to enable automated and controllable prompt generation, thereby allowing text-to-image models to produce images that can bypass AIGC detectors.

- Our method achieves strong evasion performance across both open- and closed-source detectors, and ranks first in a real-world competition. Furthermore, the generated prompts can serve as valuable adversarial datasets to support the development and evaluation of more robust and adaptive detection models.

2. Related Work

2.1. AI-generated Image Detection

Recent advancements in AIGC detection have introduced a range of effective defense strategies. Zhong et al. (Zhong et al., 2024) propose a method that suppresses global se-

mantic information and enhances local texture features by analyzing inter-pixel correlations between rich and poor texture regions, improving generalization across unseen generative models. Frank et al. (Frank et al., 2020) reveal that GAN-generated images exhibit consistent artifacts in the frequency domain caused by upsampling operations, which can be effectively exploited for detection. Liu et al. (Liu et al., 2022) focus on learning the noise patterns inherent to real images in both spatial and frequency domains, enabling the detection of generated images by identifying deviations from these natural patterns. Zhengzhe Liu et al. (Liu et al., 2020) design a detector that extracts global texture representations using Gram matrices, achieving robustness to image distortions and strong generalization to unseen fake images. Cazenavette et al. (Cazenavette et al., 2024) extract inversion features from a pre-trained diffusion model by reconstructing latent noise maps and denoised outputs, allowing their detector to generalize well to high-fidelity text-to-image models. Luo et al. (Luo et al., 2024) propose a lightweight approach that calculates latent-space reconstruction error in a single denoising step and refines image features based on spatial correlations with this error, improving both detection accuracy and efficiency.

2.2. Datasets for AI-generated Image Detection

CNNSpot(Wang et al., 2020b) collects images generated by 11 CNN-based models (e.g., ProGAN(Kang et al., 2023)) to build a dataset that covers diverse architectures and tasks, enabling evaluation of the detector's cross-generator generalization ability. GenImage(Zhu et al., 2023) adopts a unified prompt template `"photo of {class}"`, where `{class}` represents one of the 1000 categories from ImageNet(Deng et al., 2009), to generate class-controlled images. WildFake(Hong & Zhang, 2024) collects real user-generated prompts from community platforms such as Civitai(Civitai, 2025) and Midjourney(Holub, 2025), and supplements them with its own generation pipeline. The prompts cover various styles, including stylized, personalized (e.g., DreamBooth(Ruiz et al., 2022), LoRA(Hu et al., 2021)), and structure-controlled (e.g., ControlNet(Zhang et al., 2023)) types. While prior work(Papa et al., 2023) has investigated prompt-based generation of realistic faces using Stable Diffusion, its reliance on human evaluation limits the possibility of conducting systematic, model-based assessments of detection performance.

2.3. Black-box Adversarial Attack Methods

Black-box adversarial attack means that the attacker can not access the internal structure of the model and can only optimize the attack method by querying the model's output. Some methods based on random pixel updates (Vo et al., 2024; Andriushchenko et al., 2020) are effective in evading detection. Another attack method(Meng et al., 2023)
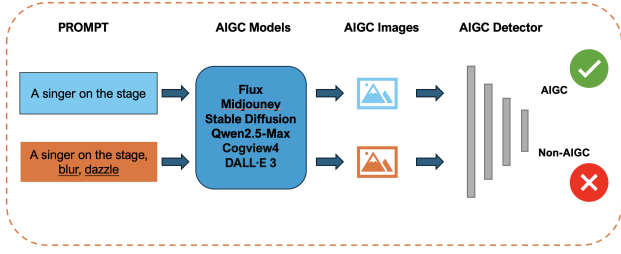
Figure 1. The adversary uses a lighting-based attack to generate images that can evade the detector.

perturbs the semantics in the attribute space of images to deceive detection models. Prior work(Xie et al., 2024) uses real-world post-processing, i.e.,Gaussian blur, JPEG compression, Gaussian noise and light spot to generate adversarial examples.

3. Threat Model & Motivation

3.1. Threat Model

We consider an adversary who aims to generate portrait images capable of bypassing AIGC detectors by crafting carefully designed text prompts for T2I models such as Flux, Midjourney, and Stable Diffusion. We focus exclusively on images generated directly by the T2I model, meaning that any form of post-processing is out of scope. The adversary has black-box access to the AIGC detectors, they can submit images and obtain detection scores, but do not know the detector's architecture.

The generated images must satisfy several predefined constraints:(1) facial features must be proportionally accurate; (2) expressions should appear natural; (3) the image should conform to basic principles of physical plausibility; (4) scenes must align with common real-world experiences and social norms.

As shown in Figure 1, a simple prompt such as *"A singer on the stage"* generates an image that is detectable as AI-generated. However, by adding a simple constrain "blur, dazzle" in the prompt, the adversary can manipulate the image's brightness, leading it to be classified as a non-AIGC image.

3.2. Prompt-based Texture Injection Attack

PatchCraft(Zhong et al., 2024) is a recently proposed framework for detecting AI-generated images. The core of PatchCraft lies in the assumption that although some cutting-edge generative models create impressive images from the semantic view, it is still hard to simulate the inter-pixel correlation of real images. Pixels in rich texture regions

exhibit more significant fluctuations than those in poor texture regions, so it proposes a Smash&Reconstruction pre-processing method to extract rich and poor texture regions from the image independently, leveraging the inter-pixel correlation contrast between rich and poor texture regions of an image as a fingerprint to identify AI-generated images.

While this method leads to strong generalization across unseen generators, it also exposes a critical vulnerability: if an attacker can manipulate the distribution or content of rich-texture regions, the detector will focus only on the injected rich-texture regions designed by the attacker, making it easy to mislead the detector.

To exploit this weakness, we design an adversarial attack against PatchCraft detector by injecting carefully crafted rich-texture regions into the generated image via prompts. Specifically, we use prompts of the form:

```
<A description of a person>
+ "with the rest of the page
filled entirely with clear
text."
```

The principle of the attack is shown in Figure 2, the description prompt ensures that the T2I model still produces a facial region as the primary content, while the "with the rest of the page filled entirely with clear text" segment guides the model to fill rest areas with densely structured textual content. Through simple testing, we found that AI-generated images composed of clear text and blank space are difficult for AIGC detectors to identify as AI-generated. These textual regions exhibit abrupt pixel variations in the image space, characteristics typical of high-frequency signals, which can be extracted by PatchCraft as rich-texture regions.

By injecting enough patches of artificial text, we effectively dominate the reconstructed rich-texture regions with adversarial content. According to the original intention of the method, the facial regions of a portrait should be extracted as rich-texture regions, but after the Smash & Reconstruction process, the detector focuses on the injected texture regions rather than the meaningful fingerprint from the face regions. Meanwhile, the poor-texture regions mostly consist of blank or low-information areas, further limiting the ability of detector to make reliable decisions.

Essentially, our attack disrupts the fundamental texture distribution that PatchCraft relies on. The classifier is misled to learn from manipulated features originating from adversarial prompts rather than natural generation artifacts, thereby inducing systematic detection failures.

3.3. Prompt-based lighting attack

In recent years, researchers have increasingly recognized that visual variations caused by lighting in natural images

Figure 2. Generated by the flux-schnell model using the prompt: `<A portrait of a person>` + "with the rest of the page filled entirely with clear text."

Figure 3. Generated using the wanx2.0-t2i-turbo model with the prompt: 'Jay Chou's live concert, clear facial features, dazzle.' The Zhuque AIGC Detector estimates a 24.3% probability that the image is AI-generated.

can serve as effective adversarial perturbations to mislead deep neural networks (Tian et al., 2021; Jiang et al., 2023; Xie et al., 2024). Unlike traditional adversarial examples that often rely on high-frequency noise, such light-based perturbations are typically more perceptually natural and harder to detect. Light-based perturbations have emerged as a natural and imperceptible means to evade deep learning models.

Early work (Tian et al., 2021) introduced adversarial vignetting attacks, simulating radial brightness falloff based on physical camera parameters. Subsequent research (Jiang et al., 2023) revealed that AIGC detectors are vulnerable to global brightness and contrast shifts. Even simple linear transformations were sufficient to mislead models, indicating a lack of robustness to natural lighting variations. Building on this, R2BA (Xie et al., 2024) proposed using localized light spots to simulate overexposure effects.

Since we are unable to apply direct post-processing to generate images in our attack scenario, we explore an alternative approach by manipulating lighting conditions through prompt engineering. Specifically, we aim to induce light-based perturbations such as dazzle or overexposure. By incorporating semantically relevant terms (e.g., *"dazzle"*) into the text prompts fed to the T2I model, this action will substantially alter the intensity distribution of pixels within the image, thereby influencing the detector's decision-making process.

Inspired by prior work on adversarial vignetting and light spot attacks, which demonstrate that changes in lighting can

impair detector performance, we aim to replicate similar effects at the T2I model generation stage using text prompts. To assess this, we evaluate the impact of prompt-induced lighting perturbations on the Zhuque AIGC detector. Our experiments reveal that incorporating light-related words into prompts (e.g., *"dazzle"*) can significantly reduce detection accuracy, indicating that the detector is sensitive to such light-related semantic cues.

Figure 3 presents a example in which we add the word "dazzle" to the prompt to control the image's lighting. This manipulation successfully induces a perturbation attack against the AIGC detector.

4. Method

4.1. Overview

As discussed in the case study in Section 3, T2I models can generate images that perturb AIGC detectors by controlling image semantic features such as texture and lighting. However, manually constructing prompts to generate images is both costly and inefficient, and it is difficult to find the optimal prompt.

To address this challenge, we developed a grammar tree-based automated prompt generator. The root node of the grammar tree is the prompt, and under the root node are several child nodes, such as "Portrait attributes" and "Stylistic attributes". The prompt generator traverses the grammar tree recursively from top to bottom to generate the prompt. We will introduce the design of the grammar tree in detail in Section 4.3.

To enable the prompt generator to efficiently create prompts that can guide T2I models to evade AIGC detectors, we utilized a variation of the Monte Carlo Tree Search (MCTS) algorithm, called the UCT-RAND algorithm (Zheng et al.,
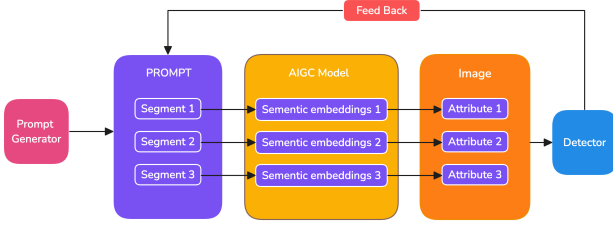
Figure 4. A illustration for generating prompts designed to bypass the AIGC detector.

Figure 5. Illustration of the Grammar Tree.

2024). UCT-Rand algorithm uses weighted random selection rather than the argmax function to choose the next child node which allows us to comprehensively explore different branches of the grammar tree.

4.2. Definition of Attribute and Semantic Embeddings

Previous research (Meng et al., 2023) on attacking Deep-Fake detectors involved introducing attribute-level perturbations to change the semantics of attributes such as mouth, hairstyle, and eyebrows. In our work, we also use attributes to represent the semantic features in the images, and we leverage semantic embeddings to represent these attributes within the T2I model's embedding space.

In this paper, we consider two types of attributes: portrait attributes and stylistic attributes. **Portrait attributes** represent human-interpretable portrait semantic features such as gender, age, and skin tone. **Stylistic attributes** include properties such as lighting and texture. For example, a T2I model can generate a image with specific lighting (the attribute), such as "overexposure" (the semantics).

The attack process can be summarized as: carefully crafted segments within the prompt can further influence the semantic embeddings in the T2I model, thereby achieving control over the attributes of the generated image, as shown in Figure 4.

4.3. Grammar Tree Design

We believe that an image's attributes are hierarchical, so when designing the grammar tree, we categorize the prompt into three components as shown in Figure 5: portrait attributes, stylistic attributes, and miscellaneous attributes. Portrait attributes are primarily used to control portrait-related semantic features such as age, gender, and skin tone. Stylistic attributes focus on semantic features like lighting and texture. The miscellaneous node includes additional constraints, for instance, "well-defined facial features" ensures the generated images have clear facial structures.

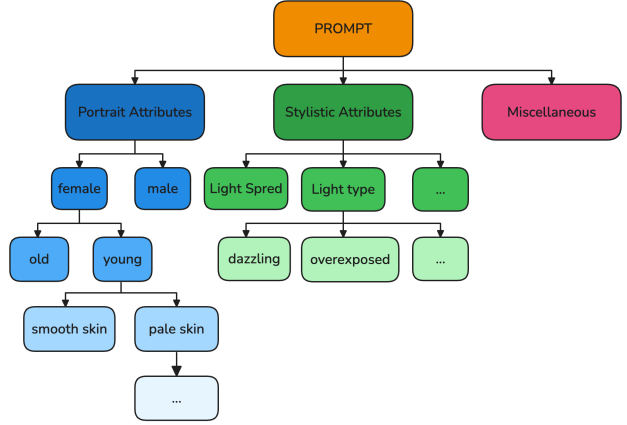There are three node types in our designed grammar tree:

RAND, OR, and AND. The AND node means that all child nodes must be visited and combined in sequence. The OR node indicates a selection among multiple candidate child nodes, where only one is chosen. The RAND node means that the number of times its child node is visited is random. If an attribute contains semantically conflicting features, we represent them using an OR node. If the features are semantically independent, we model them with an AND node. For example, the relationship between Portrait Attributes and Stylistic Attributes. When a specific semantic feature of an attribute may occur once or multiple times, we use a RAND node.

4.4. Grammar-based Prompt Generator

Prompt Generator is designed to generate high-quality prompts from a grammar tree to guide AIGC model in producing images that can evade content detectors. In this process, the Prompt Generator treats a designated non-terminal symbol (e.g., `PROMPT`) as the root node and recursively traverses the grammar tree from top to bottom. The leaf nodes serve as the terminal symbols in generated prompt.

To achieve both the efficient generation of adversarial prompts and sufficient exploration of the grammar space, we adopt the UCT-RAND algorithm (Zheng et al., 2024). The generation procedure comprises four phases: Expansion, Selection, Simulation, and Backpropagation.

Expansion. Starting from the root node, the generator traverses the grammar tree to construct a derivation tree, which ultimately forms a complete prompt. During traversal, if the current node is an AND node, all child nodes are recursively visited. For other node types, the process enters the selection phase.

Selection. If the current node is a RAND node, the number

Table 1. The number of images generated by the wanx2.0-t2i-turbo model that bypassed the detector (first 200 rounds).

ROUND	COUNT
0–49	0
50–99	3
100–149	1
150–199	1

Table 2. The number of images generated by the flux-dev model that bypassed the detector (first 400 rounds).

ROUND	COUNT
0–99	51
100–199	56
200–299	59
300–399	59

of child node visits is determined randomly within a pre-defined range; If the current node is an OR node and has unvisited children, one is selected at random; If all children have been visited, the Upper Confidence Bound formula is used to compute weights for each child, and a child node is selected via weighted random sampling.

$$C(v) := \text{weighted}_{v' \in \text{Children}(v)} \left(Q(v, v') + \sqrt{\frac{2 \ln N(v)}{N(v, v')}} \right) \quad (1)$$

Let $Q(v, v')$ be the probability of generating successfully forwarded requests by the CDN after selecting sub-node v' under node v, where $N(v)$ denotes the number of times that node v has been visited, and $N(v, v')$ represents the number of times that sub-node v' has been selected under node v. In this context, $Score_i$ represents the probability that the detector classifies an image as AI-generated, where a value of 1 indicates a 100% probability of being identified as AI-generated.

$$Q(v, v') = \frac{1}{N(v, v')} \sum_{i=1}^{N(v, v')} 2 \times (1 - \text{score}_i) \quad (2)$$

Simulation Phase: The prompt generator combines the visited leaf nodes into a prompt, which is then passed to the T2I model to generate an image. The AIGC detector will check the generated image and output an AI-generated confidence score, $Score_i$.

Back propagation Phase: Based on the score given by the detector, the generation parameters $Q(v, v')$ and $N(v, v')$ associated with the visited nodes in the grammar tree are updated accordingly..

5. Evaluation

5.1. AIGC adversarial Detection Competition

We participated in an AIGC adversarial Detection Competition (Tencent, 2025) organized by Tencent, The organizer provided Tencent's self-developed AIGC detector, Zhuque (Tencent zhuque lab, 2025).

Participants were tasked with designing prompts to generate portrait images. The objective was to design prompts to generate images that bypass the Zhuque AIGC detector. A bypass was considered successful if the generated samples achieved an AI detection rate below 50%. Participants were required to use a limited number of AIGC series APIs, such as Stable Diffusion (Esser et al., 2024), Flux (Labs, 2024), and Qwen (Wang et al., 2024), for image generation. Crucially, generated images had to exhibit normal physiological structures, adhere to physical common sense, and be logically consistent with reality. Furthermore, participants were strictly required to generate images directly using the chosen model, without employing any secondary editing tools, including built-in photo enhancement, removal, or watermark tools.

We utilized our tool with the wanx2.0-t2i-turbo (Alibaba Cloud, 2025) T2I model API to generate a sufficient number of images capable of bypassing the detectors. This demonstrated that our tool can automatically generate a large quantity of compliant images that successfully bypass the Zhuque AIGC detector, thereby showcasing its effectiveness in real-world scenarios.

5.2. Testing on Open-source Detector

To demonstrate our tool's efficacy with other T2I models, we conducted experiments on both the wanx2.0-t2i-turbo (Alibaba Cloud, 2025) and flux-dev models, using PatchCraft (Zhong et al., 2024) as the detector.

We made a particularly interesting discovery: with the PatchCraft detector, most images generated by models we considered more advanced (both intuitively and by their parameter count) were classified as AI-generated by the detector, the results show in Table 1. In contrast, weaker models like flux-dev were able to produce a significant number of images that bypassed the detector even within the first 100 iterations.

We analyzed this phenomenon and hypothesize that it's due to the PatchCraft detector's limited training data. We believe PatchCraft's heavy reliance on traditional GAN-generated datasets during training has resulted in insufficient generalization capability for detecting images from newer T2I models. Moreover, most of the training datasets for

the model are in 256×256 resolution, resulting in unstable detection performance on high-resolution images.

As Table 2 illustrates, our prompt generator can consistently produce images that bypass detectors. Furthermore, as the number of rounds increases, the proportion of prompts generated that achieve the desired target also rises.

6. Conclusion

In this paper, we present a novel adversarial prompt generation framework for T2I synthesis that effectively deceives state-of-the-art AIGC detectors. Addressing the challenges of limited controllability and the inefficiency of manual prompt engineering, we introduce a grammar tree-based generator combined with the UCT-Rand algorithm to explore the prompt space efficiently. Our method automates the creation of diverse, semantically rich prompts and significantly improves evasion success across multiple generative models. Extensive experiments show strong attack performance against both open-source and commercial detectors, revealing vulnerabilities to semantic-level prompt perturbations. In addition, our tool can be used to generate more diverse datasets for robustness training of AIGC detectors, and to support the improved design of AIGC detection frameworks.

References

Alibaba Cloud. Tongyi Wanxiang 2.1 Text-to-Image V2 API Reference. https://help.aliyun.com/zh/model-studio/text-to-image-v2-api-reference, 2025. Last updated: 2025-05-22.

Andriushchenko, M., Croce, F., Flammarion, N., and Hein, M. Square attack: a query-efficient black-box adversarial attack via random search, 2020. URL https://arxiv.org/abs/1912.00049.

Cazenavette, G., Sud, A., Leung, T., and Usman, B. 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*, pp. 10759–10769, 2024.

Chen, W., Zhang, L., Zhang, J., and Hu, H. Prompt consistency for detecting text-to-image generated content. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Chesney, R. and Citron, D. K. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs*, 98(1):147–154, 2022.

Civitai. Civitai: Ai art model sharing platform, 2025. URL https://civitai.com/. Accessed: 2025-05-18.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Douglas, J. and Yampolskiy, R. V. Deepfakes and face recognition: Challenges and opportunities. In *2021 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 448–452, 2021.

Durall, R., Keuper, M., and Keuper, J. Watch your upconvolution: 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 (CVPR)*, pp. 7890–7899, 2020.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Frank, J., Eisenhofer, T., Czempiel, T., Knörlein, B., Rigoll, G., Hilsmann, A., and Eisert, P. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning (ICML)*, 2020.

Holub, O. Midjourney, 2025. URL https://www.midjourney.com/home/. Accessed: 2025-05-18.

Hong, Y. and Zhang, J. Wildfake: A large-scale challenging dataset for ai-generated images detection, 2024. URL https://arxiv.org/abs/2402.11843.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Jiang, Z., Zhang, J., and Gong, N. Z. Evading watermark based detection of ai-generated content, 2023. URL https://arxiv.org/abs/2305.03807.

Kang, M., Zhu, J.-Y., Zhang, R., Park, J., Shechtman, E., Paris, S., and Park, T. Scaling up gans for text-to-image synthesis, 2023. URL https://arxiv.org/abs/2303.05511.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

Labs, B. F. Flux. https://github.com/black-forest-labs/flux, 2024.

Liang, C., Xu, C., Zhu, X., and Wang, L. Mindeye: Detecting ai-generated images with prompt-image alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

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

Liu, Z., Qi, X., and Torr, P. H. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8060–8069, 2020.

Luo, Y., Du, J., Yan, K., and Ding, S. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17006–17015, 2024.

Meng, X., Wang, L., Guo, S., Ju, L., and Zhao, Q. Ava: Inconspicuous attribute variation-based adversarial attack bypassing deepfake detection, 2023. URL https://arxiv.org/abs/2312.08675.

Mirsky, Y. and Lee, W. Creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., et al. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Papa, L., Faiella, L., Corvitto, L., Maiano, L., and Amerini, I. On the use of stable diffusion for creating realistic faces: from generation to detection. In *2023 11th International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–6, 2023. doi: 10.1109/IWBF57495.2023.10156981.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

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

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. 2022.

Saharia, C., Chang, W., Ho, J., Lee, C.-C., Salimans, T., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image

diffusion models with deep language understanding. In *International Conference on Machine Learning*, pp. 18235–18255. PMLR, 2022.

Tencent. Tencent zhuque lab aigc detection competition. https://tencent.zhuqueai.top/index.html, 2025. Accessed: 2025-05-19.

Tencent zhuque lab. Zhuque ai assistant: Ai-generated content detection tool. https://matrix.tencent.com/ai-detect/ai_gen, 2025. Accessed: 2025-05-26.

Tian, B., Juefei-Xu, F., Guo, Q., Xie, X., Li, X., and Liu, Y. Ava: Adversarial vignetting attack against visual recognition, 2021. URL https://arxiv.org/abs/2105.05558.

Verdoliva, L. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14 (5):910–932, 2020.

Vo, V. Q., Abbasnejad, E., and Ranasinghe, D. C. Brusleattack: A query-efficient score-based black-box sparse adversarial attack, 2024. URL https://arxiv.org/abs/2404.05311.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8695–8704, 2020a.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. Cnn-generated images are surprisingly easy to spot... for now, 2020b. URL https://arxiv.org/abs/1912.11035.

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

Xie, C., Ye, D., Zhang, Y., Tang, L., Lv, Y., Deng, J., and Song, J. Take fake as real: Realistic-like robust black-box adversarial attack to evade aigc detection, 2024. URL https://arxiv.org/abs/2412.06727.

Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1316–1324, 2018.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 5907–5915, 2017.

Zhang, H., McDermott, K., Tourazin, M., Rudd, E., and Boult, T. E. Detecting gan-synthesized faces using landmark locations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models, 2023.

Zheng, L., Li, X., Wang, C., Guo, R., Duan, H., Chen, J., Zhang, C., and Shen, K. Reqsminer: Automated discovery of cdn forwarding request inconsistencies and dos attacks with grammar-based fuzzing. In *Proceedings of the 31st Annual Network and Distributed System Security Symposium, NDSS*, volume 24, 2024.

Zhong, N., Xu, Y., Li, S., Qian, Z., and Zhang, X. Patchcraft: Exploring texture patch for efficient ai-generated image detection, 2024. URL https://arxiv.org/abs/2311.12397.

Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., and Wang, Y. Genimage: A million-scale benchmark for detecting ai-generated image, 2023.