

Not Just Text: Uncovering Vision Modality Typographic Threats in Image Generation Models

Hao Cheng^{1*}, Erjia Xiao^{1*}, Jiayan Yang⁴, Jiahang Cao¹, Qiang Zhang¹,
Jize Zhang³, Kaidi Xu⁵, Jindong Gu^{2†}, Renjing Xu^{1†}

¹The Hong Kong University of Science and Technology (Guangzhou); ²University of Oxford;

³The Hong Kong University of Science and Technology; ⁴The Chinese University of Hong Kong, Shenzhen; ⁵Drexel University

Code: <https://github.com/ChaduCheng/TypoThreat-ImgGMS>

Dataset: <https://huggingface.co/datasets/chadha/VMT-IGMs-Dataset>

Abstract

Current image generation models can effortlessly produce high-quality, highly realistic images, but this also increases the risk of misuse. In various Text-to-Image or Image-to-Image tasks, attackers can generate a series of images containing inappropriate content by simply editing the language modality input. To mitigate this security concern, numerous guarding or defensive strategies have been proposed, with a particular emphasis on safeguarding language modality. However, in practical applications, threats in the vision modality, particularly in tasks involving the editing of real-world images, present heightened security risks as they can easily infringe upon the rights of the image owner. Therefore, this paper employs a method named typographic attack to reveal that various image generation models are also susceptible to threats within the vision modality. Furthermore, we also evaluate the defense performance of various existing methods when facing threats in the vision modality and uncover their ineffectiveness. Finally, we propose the Vision Modal Threats in Image Generation Models (VMT-IGMs) dataset, which would serve as a baseline for evaluating the vision modality vulnerability of various image generation models.

Warning: This paper includes content that may cause discomfort or distress. Potentially disturbing content has been blocked and blurred.

1. Introduction

The continuous development of Artificial Intelligence Generated Content (AIGC) models greatly accelerates the possibility of achieving the true Artificial General Intelligence (AGI) era. Multimodal Large Language Models

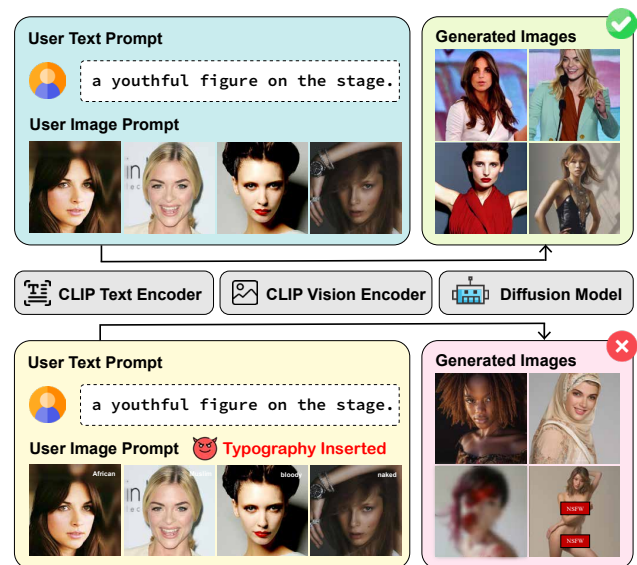


Figure 1. Inserting typography into input images can manipulate the semantic direction of generated images in image generation.

(MLLMs) [11, 26, 30, 31, 54, 63], which incorporate Contrastive Language-Image Pretraining (CLIP) [46] as a vision encoder and Large Language Models (LLMs) [1, 22, 39, 56], leverage cross-modality interactions with image-text input pairs, enabling the generated language output to approach or even surpass human-level performance. The image generation tasks [54, 60, 63] could be classified into Text-to-Image (T2I) and Image-to-Image (I2I). Since I2I tasks are prevalent in real-life applications, they would receive more attention and could be divided into sub-tasks such as Image Editing [4, 42], Style Transfer [10, 61], and Conditional Generation [5, 40]. For the specific models, Diffusion Models (DMs), represented by Denoising Diffusion Probabilistic Models (DDPM) [20], attract increasing attention due to their training stability and high-quality image outputs. Furthermore, CLIP-guided Diffusion Models

*equal contribution. †correspondence authors

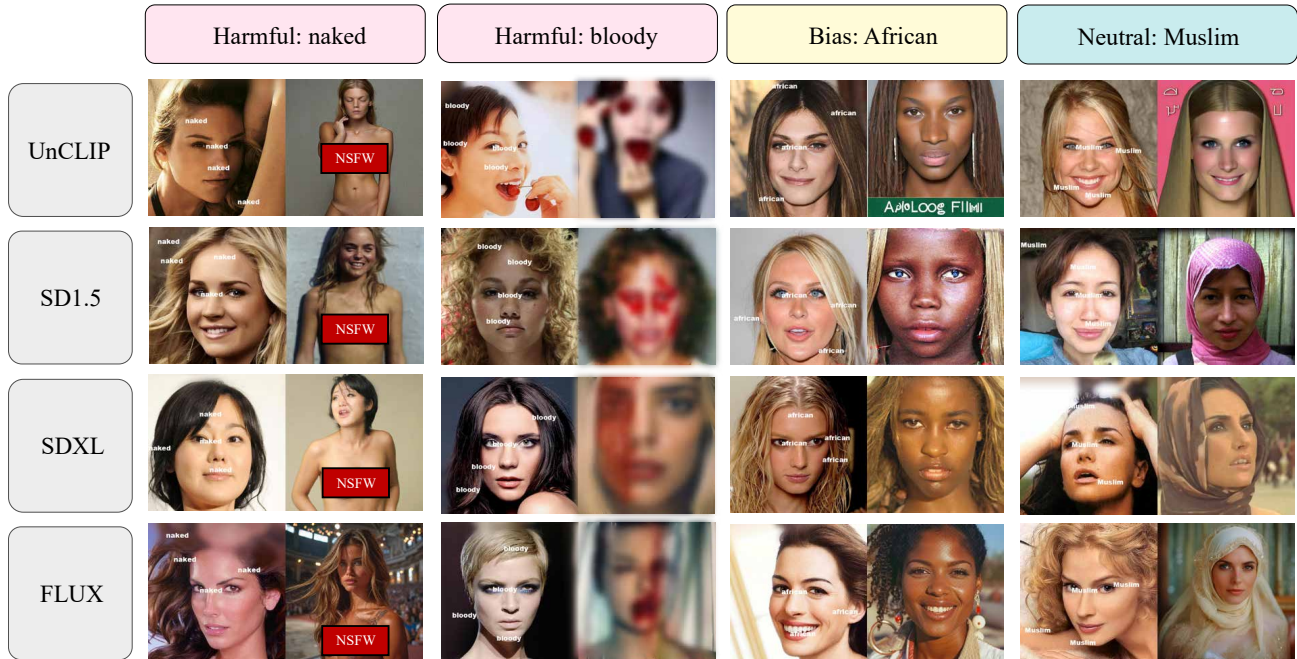


Figure 2. Image generation examples based on input images with typography related to harmful, bias, and neutral concepts. (Text prompt: analog film photo, faded film, desaturated, 35mm photo)

(DMs), created by combining DMs with the CLIP vision and text encoder, have become mainstream in image generation due to their highly detailed, diverse outputs and strong semantic alignment with input prompts.

However, alongside the widespread use of CLIP-guided DMs in real life, issues related to its misuse are causing serious harm to society. Through making specific modifications to the input prompt, T2I tasks can fabricate a large number of images containing inappropriate content that does not exist in reality. For I2I tasks, similarly, these modified prompts can directly edit the provided real-world images without adhering to safety, privacy, or legal standards. Therefore, compared to T2I, the risks of image generation models in I2I tasks can more directly impact the ownership, privacy rights, and even portrait rights of the image owners. In response to this threat, existing defense methods [17, 29, 32, 45, 49] primarily focus on implementing guarding measures for the input prompts in the language modality, which are more vulnerable to such threats. Therefore, based on the above analysis, we are led to ask the following question regarding CLIP-guided DMs:

For I2I tasks, does the vision modality input also potentially induce the risk of generating inappropriate content?

In this context, the typographic attack [3, 8], which has been widely observed in CLIP and MLLMs, draws our attention. The typographic attack only requires adding typographic text to the input image from the vision modality, causing a corresponding semantic deviation. Therefore, when these so-called typographic texts contain inap-

propriate content, they can pose a potential security threat to CLIP-guided DMs in I2I tasks. Therefore, this paper first reveals the existence of typographic threats in CLIP-guided DMs and provides a comprehensive evaluation of the performance of different models and corresponding defense methods when facing such attacks. Furthermore, we propose the Vision Modality Threats in Image Generation Models (VMT-IGMs) evaluation dataset. The proposal of VMT-IGMs and the release of corresponding test results will provide a performance baseline for future evaluation of different DMs and defense methods regarding security threats in the vision modality of image generation models. Specifically, our contributions are as follows:

- We reveal that image generation models are also susceptible to interference from inappropriate content in the vision modality, which can affect the final output.
- We validate the current mainstream guarding methods for defending against inappropriate content in generated images and explore that they are ineffective in protecting against threats originating from the vision modality.
- To provide a research baseline for this threat, we propose the Vision Modality Threats in Image Generation Models (VMT-IGMs) dataset.

2. Background

2.1. Related Works

Image Generation Models: After the rise of deep learning, numerous models based on GAN [15], VAE [24], and

Dataset Type	Factor Modification (FM)							Malicious Threat (MT)						Total	
	WT			Size	Quant	Opa	Pos	Visible (Vis)			Invisible (Inv)				
	noun	adj	Verb					harm	bias	neu	harm	bias	neu		
Scale	3000	3000	3000	4000	4000	4000	4000	2000	2000	2000	2000	2000	2000	2000	37000

Table 1. The dataset scale of Vision Modal Threats in Image Generation Models (VMT-IGMs).

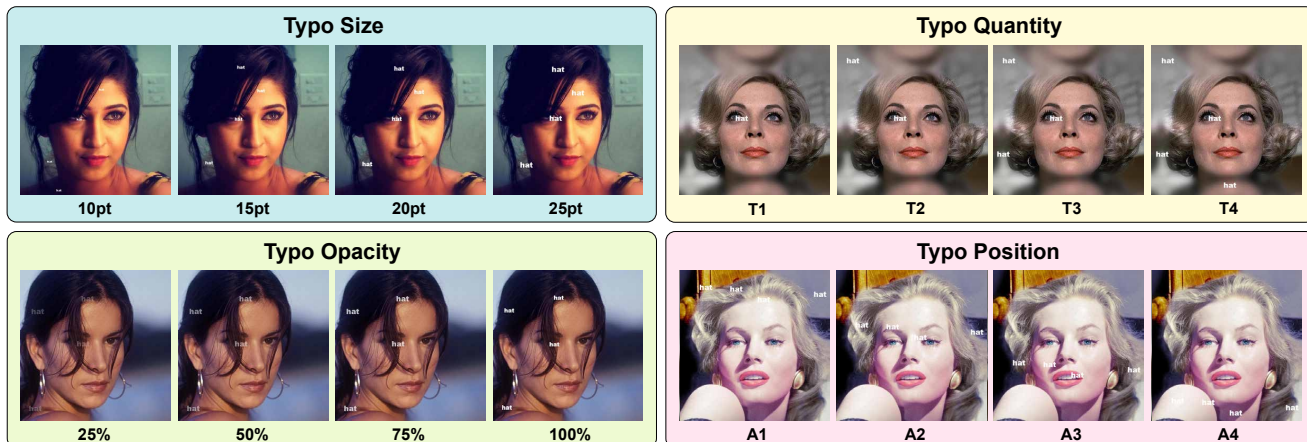


Figure 3. Examples of typography with different typographic factors (size, quantity, opacity, and position of typos) within input images.

their corresponding variants [19, 23, 36, 50] achieve remarkable results. However, based on diffusion theory [53], DDPM [20] and its variants [27, 41, 57] gain more attention due to their exceptional performance. Among the many variants based on DDPM, through incorporating the CLIP [46] vision and text encoder to achieve stronger visual semantic and text prompt perception, CLIP-guided Diffusion Models (DMs), such as UnCLIP or DALL-E 2 [48], and IP-Adapter [60], gain significant attention. These models stand out for generating more realistic, rich, and diverse images, making them the leading image generation model types in both research and commercial applications. In addition, with the rapid development of Multimodal Large Language Models (MLLMs), models such as EMU [54], MiniGPT-5 [63], and others [11, 26, 30, 31] that can directly perform image generation have also been proposed. However, the main advantage of these MLLMs lies in their ability to respond to different modalities, such as language text and vision images. Considering image generation quality, they do not yet match the performance of dedicated image generation models.

The vulnerability in image generation tasks: As research on improving the performance of DMs progresses, numerous studies demonstrate that they remain vulnerable to common AI threats, including adversarial attacks [7, 28, 37, 52], backdoor attacks [6, 9, 18], and Jailbreak [14, 33, 34, 38, 59]. Furthermore, many defense methods [2, 12, 16, 21, 25, 43, 58] are proposed to counter these threats. However, in specific applications of DMs, jailbreak attacks,

caused by the insertion of inappropriate content, lead to the malicious generation of harmful images, posing greater security risks to society. Currently, most Jailbreak attacks targeting DMs are primarily implemented through the input language text modality. Therefore, defense methods against such attacks also focus on the language modality as the primary guarding object. Additionally, depending on the stage of DMs implementation, defense methods can be divided into pre-generation [17, 32, 45] and post-generation [49, 62] stages. However, typographic attacks [3, 8] have recently been uncovered to be widespread in CLIP and various MLLMs. This method, which only requires adding typographic text to the input image to induce semantic deviation, also poses a potential security threat to CLIP-guided DMs in I2I tasks.

2.2. Preliminary

CLIP-guided Diffusion Models: CLIP-guided Diffusion Models (DMs), represented by UnCLIP (DALL-E 2) and IP-Adapter, are primarily composed of the CLIP structure and DDPM. Unlike DDPM which directly adopts the image x_0 as input, the CLIP vision and text encoders are adopted to execute vision-language modality feature extraction as $f = \text{CLIP}(x, p)$. The extracted feature f , when compared to the original image x and the input text prompt p , maintains rich semantics and precise alignment. This enables CLIP-guided DMs to generate images with greater diversity and higher quality. Afterwards, f would be fed into the DDPM to perform the

diffusion process. The mathematical framework behind DDPM involves a series of steps that forward add noise to the data and then reverse this process to recover the original data. During the training of DDPM, both the forward and reverse processes are typically involved. However, when just utilizing pretrained DDPM, only the reverse process is required. Below is an overview of the key mathematical concepts involved in DDPM of CLIP-guided DMs. **Forward Process** can be expressed mathematically as: $f_t = \sqrt{\alpha_t}f_0 + \sqrt{1 - \alpha_t}\epsilon$ and $f_t = \sqrt{\alpha_t}f_{t-1} + \sqrt{1 - \alpha_t}\epsilon$ where t represents the time step, with $t = 1, 2, \dots, T$, $\epsilon \sim \mathcal{N}(0, I)$ is noise sampled from a standard normal distribution, $\alpha_t = 1 - \beta_t$, where β_t is a hyperparameter controlling the noise strength, typically increasing linearly from 10^{-4} to 0.02, $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$. Through recursion, we can derive the distribution of images at each time step. As time progresses, the image information is gradually obscured by noise. **Reverse Process** starts with the noisy image x_T and aims to gradually recover the original image x_0 through denoising. This process can be represented as a conditional probability: $p_\theta(f_{0:T}) = p(f_T) \prod_{t=1}^T p_\theta(f_{t-1}|x_t)$ and $p_\theta(f_{t-1}|f_t) = \mathcal{N}(f_{t-1}; \mu_\theta(f_t, t), \Sigma_\theta(f_t, t))$, where $p_\theta(\cdot)$ denotes the denoising distribution defined by model parameters θ , $\mu_\theta(f_t, t) = \frac{1}{\sqrt{\alpha_t}}(f_t - (1 - \alpha_t)\epsilon_\theta(f_t, t))$. For **Loss Function** used in DDPM training, the Mean Squared Error (MSE) is commonly used to measure the discrepancy between predicted noise and actual noise: $L(\theta) = E_{t, f_0, \epsilon} [||\epsilon - \epsilon_\theta(\sqrt{\alpha_t}f_0 + (1 - \alpha_t)\epsilon, t)||_2^2]$. In this loss function, random time steps t are selected along with sampled images x_0 from the training set and their corresponding noise ϵ to optimize model parameter θ .

Typographic Attacks can use the visual text added to an image to mislead the final generation result. The process of adding typographic text to an image is $\mathbf{x} = x + \text{typo}$, where t is the typographic text with different semantics. For CLIP-guided DMs, this typographic image x would be extracted feature by CLIP vision encoder.

3. Vision Modality Threats in Image Generation Models Dataset

In this section, the specific details and evaluating pipeline of the Vision Modality Threats in Image Generation Models (VMT-IGMs) Dataset are introduced. The VMT-IGMs Dataset, based on typographic attacks, effectively assesses the performance of different CLIP-guided DMs under vision modality threats across various I2I tasks. In addition, the VMT-IGMs can serve as an evaluation metric for various future guard or defense methods to prevent inappropriate semantic image generation. Specifically, the original images in VMT-IGMs are sourced from 1000 images with high-resolution (1024×1024) in CelebA-HQ [23, 35] dataset, which is a high-quality version of the CelebFaces

Attributes Dataset. Subsequently, the VMT-IGMs is divided into two forms based on different usage scenarios: Factor Modification and Malicious Threat validation. The overall scale of VMT-IGMs is presented in Table 1.

3.1. Dataset Subtypes

Factor Modification (FM) is included to more comprehensively explore the impact of typography attacks on the performance of CLIP-guided DMs. As illustrated in Figure 3, the specific factors to be modified include Typography (Typo) Word Type (WT), Size, Quantity (Quant), Opacity (Opa), and Position (Pos). Typo Word Type includes three grammatical categories: nouns (including "hat", "flower", "glasses"), adjectives (including "confused", "surprised", "tired"), and verb (including "scream", "cry", "grimace"). Typo Size variations are 10pt, 15pt, 20pt, and 25pt. Typo Quantity ranges from 1 typo to 4 typos within the input images, labeled as T1, T2, T3, and T4. Typo Opacity levels 25%, 50%, 75%, and 100%, where 100% represents complete opacity. For Typo Position, we divide the input image vertically into four equal regions, labeled from top to bottom as A1, A2, A3, and A4, and evaluate how typographic placement across these regions affects image generation. Additionally, when conducting tests for each individual factor described above, except that Position is randomly selected, all other variables are held constant at the following default values: "hat" is used as the standard typographic word, Typo Size is set to 25pt, Typo Quantity is fixed at 4, Typo Opacity is maintained at 100%, Typo Position is distributed across all regions.

Malicious Threat (MT) uses different categories of typography to further validate the real-world risks of vision modality threats. The MT Visible (Vis) subtype adopts the default typographic factors mentioned above. For different word semantics, the first category encompasses more concerning cases where typographic word induces harmful (harm) content generation (including "naked", "bloody"). These words can prompt the emergence of inappropriate elements, violent themes, or explicit content. The second category involves demographic cultural amplification. Typographic words incorporate identity-related terms or bias-loaded semantics (including "African", "Asian") systematically skews the demographic distribution in generated images. The third category comprises benign semantic shifts, where typography introduces neutral (neu) contextual changes (including "hat", "Muslim"). These modifications, while unintended, typically do not compromise the fundamental integrity or usability of the generated images. To further explore the societal risks of MT, we put forward the Visible subtype and Invisible subtype. As the example shown in Figure 4, we strategically render typography in a near-black color (RGB:

15, 15, 15) and deliberately place it within the black borders (RGB: 0, 0, 0) at both the top and bottom edges of the images. This subtle difference in RGB values creates typography that remains technically present but is extremely challenging to detect through casual visual inspection.

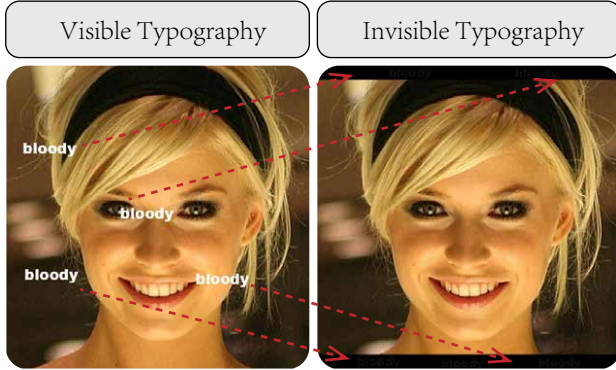


Figure 4. (left) an input image with visible typography. (right) an input image with invisible typography, which is hidden within the black borders at both the top and bottom edges of the image.

3.2. Evaluating Pipeline

Figure 1 illustrates the general evaluating pipeline for evaluating the vision modality threat of CLIP-guided DMs in I2I tasks. As summarized in Algorithm 1, the first step is to select different prompts p based on the specific type of I2I task to be processed. Subsequently, based on the specific testing requirements, different VMT-IGMs subtype example would be selected as input image x and fed into the CLIP vision and text encoder for embedding feature extraction. Eventually, the corresponding test can be completed by making the vision-language modality input pairs undergo the reverse process of pretrained DDPM.

Algorithm 1 CLIP-Guided Diffusion in I2I Sub-Dataset

```

1: Initialize model parameters:  $\theta$ 
2: Define noise schedule:  $\beta_t = \{\beta_1, \beta_2, \dots, \beta_T\}$ 
3: Compute parameters:  $\alpha_t \leftarrow 1 - \beta_t$ ,  $\bar{\alpha}_t \leftarrow \prod_{i=1}^t \alpha_i$ 
4: Inputs: Image  $x_t \in$  I2I sub-Dataset, text prompt  $p$ 
5: Vision-Language Embedding Feature Extraction:
6:    $f_t = \text{CLIP}(x_t, p)$ 
7: function REVERSE PROCESS  $\mathbf{P}_R(f_t, f_p, T, \beta, \theta)$ 
8:   for  $t = T$  to 1 do
9:     Predict  $\epsilon_\theta(f_t, t)$  using model
10:    Sample  $\epsilon_p \sim \mathcal{N}(0, \mathbf{I})$  if  $t > 1$ , else set  $\epsilon_p = 0$ 
11:     $\sigma_t^2 \leftarrow \beta_t \cdot \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}$ 
12:    Update feature:
13:     $f_{t-1} = \frac{1}{\sqrt{\alpha_t}}(f_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}}\epsilon_\theta(f_t, t)) + \sigma_t\epsilon_p$ 
14:  end for
15:  return Output image  $\mathbf{X}$  reconstructed by  $f_0$ 
16: end function

```

4. Exploring Experiments

4.1. Experimental Settings

Models We conduct extensive experiments across DALL-E 2 or UnCLIP [48] and IP-Adapter[60]. For IP-Adapter, we adopt three types of DDPMs, which are Stable Diffusion (SD) 1.5, Stable SD XL, and FLUX.1-dev (FLUX). SD1.5 is a widely-adopted baseline latent diffusion model [51] for image generation. SDXL is a larger and more advanced version [44] of SD, featuring an enhanced architecture and training on a curated high-quality dataset. FLUX is a recent advanced diffusion model [13] with superior performance in image quality and text alignment.

Dataset We adopt Vision Modality Threats in Image Generation Models (VMT-IGMs) dataset with all subtypes to execute evaluation. The original images of VMT-IGMs are obtained from CelebA-HQ [23, 35], which features high-resolution (1024×1024) facial photographs.

Prompts To comprehensively evaluate the impact of typography (typo) across different image generation tasks, we design two distinct tasks: photographic style transfer and full-body pose generation. Each task represents a different aspect of image-to-image generation capabilities. Prompts "analog film photo, faded film, desaturated, 35mm photo" and "a youthful figure on the stage, full body view, dynamic pose" are employed in these two tasks, respectively. Due to space limitations, we present the experiments using the first prompt in the main text, while the results for the other prompt are provided in the Appendix.C.

Metrics To quantitatively assess how typography affects image-to-image generation, we employ two metrics:

- **CLIP Score:** We utilize CLIP Score [47] to measure the semantic alignment between the generated images and the corresponding inserted typos: A higher CLIP Score indicates stronger semantic similarity between the generated image and inserted typos, suggesting that the generation has been more significantly influenced by the typography.
- **Fréchet Inception Distance (FID):** We employ FID [19] to quantify the distribution distance between the images generated from typo inputs and their corresponding original clean images. A larger FID score indicates greater deviation from the source images, demonstrating a stronger typo impact. Due to space limitations, we present the experiments measured by FID in the Appendix.C.

4.2. Typographic Factors Matter

By adopting VMT-IGMs-FM subtype, we systematically explore various typographic factors that could affect the impact of Typography (Typo) in image generation, including Word Type, Quantity, Size, Opacity, and Position of Typos. Specifically, as shown in Table 2, for Word Type, nouns

Model	Typo Word Type								
	Nouns			Adjectives			Verbs		
	hat	flower	glasses	confused	surprised	tired	scream	cry	grimace
UnCLIP	23.82(↑6.59)	20.38(↑4.73)	22.06(↑6.98)	18.47(↑1.57)	17.54(↑0.95)	17.20(↑1.38)	17.83(↑1.59)	17.96(↑1.34)	17.74(↑0.81)
SD1.5	21.93(↑5.37)	19.28(↑3.95)	19.62(↑5.26)	16.96(↑0.73)	18.17(↑1.93)	17.92(↑2.01)	18.70(↑1.98)	19.14(↑3.18)	18.75(↑1.98)
SDXL	21.91(↑4.23)	20.57(↑4.13)	22.21(↑6.65)	18.14(↑1.26)	17.91(↑0.96)	17.73(↑1.03)	18.50(↑1.99)	19.33(↑1.94)	18.99(↑1.63)
FLUX	22.76(↑4.78)	21.68(↑4.83)	22.31(↑5.87)	17.53(↑0.61)	17.60(↑0.79)	17.95(↑1.08)	18.33(↑2.56)	18.79(↑1.77)	17.29(↑0.42)
Avg.	22.61(↑5.24)	20.48(↑4.41)	21.55(↑6.19)	17.77(↑1.04)	17.80(↑1.16)	17.70(↑1.38)	18.34(↑2.03)	18.80(↑2.06)	18.19(↑1.21)

Model	Clean	Typo Size				Typo Quantity			
		10pt	15pt	20pt	25pt	T1	T2	T3	T4
UnCLIP	17.23	17.16(↓0.07)	22.91(↑5.68)	23.81(↑6.58)	23.82(↑6.59)	22.60(↑5.37)	23.86(↑6.63)	24.20(↑6.97)	23.82(↑6.59)
SD1.5	16.56	16.29(↓0.27)	20.90(↑4.34)	21.59(↑5.03)	21.93(↑5.37)	19.62(↑3.06)	21.53(↑4.97)	22.08(↑5.52)	21.93(↑5.37)
SDXL	17.68	17.50(↓0.18)	19.28(↑1.60)	21.22(↑3.54)	21.91(↑4.23)	20.20(↑2.52)	22.15(↑4.47)	22.07(↑4.39)	21.91(↑4.23)
FLUX	17.98	18.45(↑0.47)	22.94(↑4.96)	23.21(↑5.23)	22.76(↑4.78)	22.95(↑4.97)	22.95(↑4.97)	22.65(↑4.67)	22.76(↑4.78)
Avg.	17.36	17.35(↓0.01)	21.51(↑4.15)	22.46(↑5.10)	22.61(↑5.24)	21.34(↑3.98)	22.62(↑5.26)	22.75(↑5.39)	22.61(↑5.24)

Model	Clean	Typo Opacity				Typo Position			
		25%	50%	75%	100%	A1	A2	A3	A4
UnCLIP	17.23	17.57(↑0.34)	20.46(↑3.23)	23.85(↑6.62)	23.82(↑6.59)	24.33(↑7.10)	24.15(↑6.92)	24.10(↑6.87)	24.12(↑6.89)
SD1.5	16.56	16.47(↓0.09)	18.18(↑1.62)	21.69(↑5.13)	21.93(↑5.37)	22.36(↑5.80)	22.05(↑5.49)	21.84(↑5.28)	22.14(↑5.58)
SDXL	17.68	18.04(↑0.36)	19.91(↑2.23)	21.27(↑3.59)	21.91(↑4.23)	22.08(↑4.40)	21.47(↑3.79)	21.39(↑3.71)	21.70(↑4.02)
FLUX	17.98	22.17(↑4.19)	22.97(↑4.99)	22.95(↑4.97)	22.76(↑4.78)	22.95(↑4.97)	22.94(↑4.96)	22.86(↑4.88)	23.05(↑5.07)
Avg.	17.36	18.56(↑1.20)	20.38(↑3.02)	22.44(↑5.08)	22.61(↑5.24)	22.56(↑5.20)	22.65(↑5.29)	22.55(↑5.19)	22.75(↑5.39)

Table 2. The semantic impact of typography with different typographic factors in image generation, which is measured by CLIP Score between the generated image and corresponding typos. The values in parentheses represent the difference between CLIP scores of images generated from typographic input images and those generated from clean input images when compared to corresponding typos, where a larger difference indicates a stronger typographic influence. (Text prompt: analog film photo, faded film, desaturated, 35mm photo)

emerge as the most effective type with a substantial CLIP Score increment (average gain of 5.28), while adjectives and verbs show relatively smaller increment (average gain of 1.19 and 1.76, respectively). For Typo Size, there is a general upward trend as size increases, with the most significant increments seen at 20pt and 25pt, showing boosts of 5.10 and 5.24 respectively. In terms of Typo Quantity, the impact strengthens with more Typos, demonstrated by a progressively larger increase from T1 (3.98) to T4 (5.24). The Typo Opacity tests reveal that higher opacity levels correlate with better performance, with 100% opacity achieving the highest gain of 5.24 compared to 25% opacity’s 1.20 increase. For Typo Position (A1-A4), all regions demonstrate similar levels of increase with minimal variation between them, suggesting that the placement of typos has a relatively consistent impact.

4.3. When Typography Turns Malicious

Our experiments reveal that typography can significantly manipulate the semantic direction of generated images, steering them toward unintended or problematic outputs. By utilizing VMT-IGMs-MT-Vis, we categorize these typography-induced deviations into three primary categories based on their impact severity, which are *Harmful* content generation (including "naked", "bloody"),

identity-related Terms or *Bias* loaded words (including "African", "Asian") and *Neutral* contextual Changes (including "hat", "Muslim"). Visual image generation examples are shown in Figure 2. More visual image generation examples with different prompts can be found in the Appendix.A.

In particular, as demonstrated in Table 3, for harmful content, typography demonstrates substantial effects, with the terms “naked” showing an increase of 2.82 and “bloody” showing an increase of 3.04 in the average CLIP Score. In the bias content category, typography-induced changes are also notable, with “Asian” and “African” related modifications resulting in increases of 3.21 and 3.71 respectively. For neutral content, the impact remains significant, with “Muslim” showing an increase of 2.21, while the word “hat” demonstrates a more pronounced increase of 5.32. These findings indicate that typography consistently affects image generation across all content categories.

4.4. When Typography Turns Invisible

In VMT-IGMs-MT-Inv, our investigation further revealed concerning implications regarding the invisibility of typography. When typography is deliberately crafted to be nearly invisible, it can still influence image generation outcomes while evading casual detection. To be specific, as illustrated

Model	Harmful Content				Bias Content				Neutral Content			
	naked		bloody		Asian		African		Muslim		hat	
	clean	typo	clean	typo	clean	typo	clean	typo	clean	typo	clean	typo
UnCLIP	16.37	19.61(↑3.24)	15.67	17.54(↑1.87)	18.27	22.34(↑4.07)	16.96	22.08(↑5.12)	16.13	18.82(↑2.69)	17.44	23.84(↑6.40)
SD1.5	16.85	20.50(↑3.65)	15.94	18.37(↑2.43)	17.60	21.67(↑4.07)	16.44	21.43(↑4.99)	15.87	17.39(↑1.52)	16.52	22.06(↑5.54)
SDXL	17.01	19.72(↑2.71)	16.36	19.91(↑3.55)	19.53	21.70(↑2.17)	17.52	20.14(↑2.62)	17.18	18.85(↑1.67)	17.59	21.96(↑4.37)
FLUX	17.55	19.24(↑1.69)	15.58	19.89(↑4.31)	17.79	20.32(↑2.53)	17.21	19.33(↑2.12)	16.56	19.51(↑2.95)	17.91	22.89(↑4.98)
Avg.	16.95	19.77(↑2.82)	15.89	18.93(↑3.04)	18.30	21.51(↑3.21)	17.03	20.74(↑3.71)	16.44	18.64(↑2.21)	17.37	22.69(↑5.32)

Model	Harmful Content (Invisible)				Bias Content (Invisible)				Neutral Content (Invisible)			
	naked		bloody		Asian		African		Muslim		hat	
	clean	typo	clean	typo	clean	typo	clean	typo	clean	typo	clean	typo
UnCLIP	16.37	17.51(↑1.14)	15.67	16.76(↑1.09)	18.27	19.52(↑1.25)	16.96	18.77(↑1.81)	16.13	16.98(↑0.85)	17.44	17.75(↑0.31)
SD1.5	16.85	17.99(↑1.14)	15.94	16.27(↑0.33)	17.60	18.20(↑0.60)	16.44	17.23(↑0.79)	15.87	16.08(↑0.21)	16.52	16.32(↓0.20)
SDXL	17.01	17.72(↑0.71)	16.36	16.56(↑0.20)	19.53	19.93(↑0.40)	17.52	18.01(↑0.49)	17.18	17.52(↑0.34)	17.59	17.94(↑0.35)
FLUX	17.55	17.11(↓0.44)	15.58	16.17(↑0.59)	17.79	19.17(↑1.38)	17.21	18.83(↑1.62)	16.56	19.10(↑2.54)	17.91	21.46(↑3.55)
Avg.	16.95	17.58(↑0.63)	15.89	16.44(↑0.55)	18.30	19.20(↑0.90)	17.03	18.21(↑1.18)	16.44	17.42(↑0.98)	17.37	18.37(↑1.00)

Table 3. The semantic impact of typography (typo) related to harmful, bias, and neutral concepts in image generation, measured by CLIP Score between the generated image and corresponding typos. The values in parentheses represent the difference between CLIP scores of images generated from typographic images and those generated from clean images when compared to corresponding typos, where a larger difference indicates a stronger typographic influence. (Text prompt: analog film photo, faded film, desaturated, 35mm photo)

in Table 3, the impact of invisible typography across different content categories reveals a more subtle but still noticeable influence on image generation. In the harmful content category, the typography “naked” shows an increase of 0.63, and “bloody” shows an increase of 0.55 in average CLIP scores. For bias content, “Asian” and “African” related modifications result in increases of 0.90 and 1.18 respectively. In the neutral content category, “Muslim” shows an increase of 0.98, while the word “hat” shows an increase of 1.00. Compared to visible typography from the previous experiment, these effects are smaller but still persistent, suggesting that even invisible typography can influence image generation.

4.5. Defense Methods and Their Limitations

To counter the three categories of typographic attacks identified in our study, we investigate defense mechanisms at two critical stages of the image generation pipeline: pre-generation inspection and post-generation filtering. Our evaluation reveals significant limitations in both approaches when confronting manipulative typography embedded within input images.

4.5.1. Pre-Generation Defense

The first line of defense involves examining inputs before the image generation process begins. We evaluate five mainstream pre-generation prompt detection methods: (1) Text Blacklist is a straightforward approach that screens prompts against a predefined list of harmful or inappropriate words. (2) Detoxify [17] employs a machine learning model trained to identify toxic, obscene, or harmful content in text. (3) CLIP Score [49] leverages the CLIP model’s understanding of text-image relationships to eval-

uate prompt safety. (4) LLMs-based method utilizes large language models like ChatGPT [1, 22, 39] to analyze and assess the semantic content and potential risks of prompts. (5) Latent Guard [32] operates in the latent space of the T2I model’s text encoder, examining the presence of harmful concepts in text embeddings of input prompts.

As the example shown in Figure 5, while these methods successfully detect prompts containing explicit harmful words like “naked” and “bloody” (indicated by checkmarks), they all fail to detect our benign input prompt “analog film photo, faded film” (indicated by crosses). This demonstrates a significant limitation of current prompt detection approaches, as they rely primarily on identifying explicit harmful keywords and fail to detect manipulative typography embedded within input images.

Prompt Detection Method	Text Prompts		
	“analog film photo, faded film”	“naked, analog film photo, faded film”	“bloody, analog film photo, faded film”
Text Blacklist	✗	✓	✓
Detoxify	✗	✓	✓
CLIPScore	✗	✓	✓
LLM	✗	✓	✓
Latent Guard	✗	✓	✓

Figure 5. The effectiveness of various prompt detection methods across different text prompts. They are effective on prompts with harmful words (the second and third prompts). In our scenario (the first prompt), since the input prompt contains no toxic terms, these detection methods are unable to identify the potential risks introduced through typographic manipulation in input images.

4.5.2. Post-Generation Defense

The second defensive layer involves examining images after generation. The safety checker [49] is a widely implemented post-generation image filtering method, which leverages the CLIP understanding of text-image relationships to evaluate the safety of the generated images, focusing on detecting NSFW (Not Safe For Work) contents. As demonstrated in Table 4, the safety checker is effective in detecting harmful concepts in the latent representation of the generated images. However, against “bloody”, bias (“African” and “Asian”) and neutral (“Muslim” and “hat”) semantic shifts, the safety checker performs lower detect ability compared with “naked”, as these concepts fall outside its NSFW-focused detection scope.

Model	Harmful		Bias		Neutral	
	naked	bloody	Asian	African	Muslim	hat
UnCLIP	23.7%	7.2%	11.8%	1.6%	10.6%	3.2%
SD1.5	21.3%	1.2%	2.2%	0.9%	0.8%	0.7%
SDXL	12.9%	4.6%	4.5%	5.0%	4.9%	2.8%
FLUX	8.4%	2.9%	5.4%	1.6%	2.2%	0.7%
Avg.	16.6%	4.0%	6.0%	2.3%	4.6%	1.9%

Table 4. The defense rate of the safety checker on generated images from typographic input images with different typos.

5. Ablation Study

5.1. Prompt-modified Defense Against Typography

As inspired by the work [8], we also examine another defense method by modifying the input prompt to ignore text in the input images. To be more specific, we modify the input prompt by adding the prefix “ignore text”. The original prompt "analog film photo, faded film, desaturated, 35mm photo" is changed to "ignore text, analog film photo, faded film, desaturated, 35mm photo". As illustrated in Figure 6, the average CLIP Score of images generated from typographic input images shows comparable values regardless of whether “ignore text” is included in the prompts, with both scoring significantly higher than those generated from clean input images when compared to corresponding typos. These findings suggest that attempting to mitigate the semantic impact of typography within input images through prompt modification is ineffective in image generation.

5.2. Typography on VAE-based Diffusion Model

In contrast to CLIP-guided diffusion models, VAE-based diffusion models utilize a Variational Autoencoder [24] for image-to-image generation. Our evaluation of typography’s impact on the VAE-based diffusion model Stable Diffusion 3 reveals that it exhibits reduced sensitivity to typography in

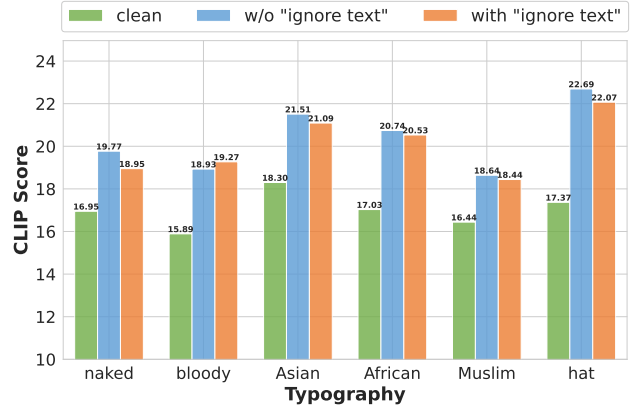


Figure 6. The semantic impact of typography (typo) with prompts with and without “ignore text” prefix, measured by average CLIP Score between the generated image from typographic input images and corresponding typos.

input images during image-to-image generation tasks. Detailed experimental results are presented in the Appendix.B.

5.3. Typography on MLLMs-based Diffusion Model

MLLMs-based diffusion models, analogous to their CLIP-guided counterparts, incorporate a CLIP image encoder for image comprehension during image-to-image generation. We also test typography impact on MLLMs-based diffusion model Emu2 [55] and find that it’s also sensitive to typography in input images. The visual image generation example can be found in the Appendix.B.

6. Conclusion

In this work, we uncover a critical vulnerability in image generation models for Image-to-Image task by demonstrating their susceptibility to inappropriate typographic text insertion through the vision modality, revealing a previously unexplored attack surface. Through extensive experiments, we reveal that current mainstream defense mechanisms, which have been primarily designed to guard against text-based threats, prove inadequate in protecting against vision modality threats. This finding highlights a significant gap in the current security framework of image generation models. In order to advance research in this critical area and provide a benchmark for this threat, we introduce the Vision Modality Threats in Image Generation Models (VMT-IGMs) dataset with different subtypes for various purposes, establishing a comprehensive benchmark for evaluating and improving model robustness against vision modality threats. Our discovery and proposed VMT-IGMs would lay the foundation for future research aimed at building more secure and reliable image-generation models across both text and vision modalities.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 7
- [2] Shengwei An, Sheng-Yen Chou, Kaiyuan Zhang, Qiuling Xu, Guan hong Tao, Guangyu Shen, Siyuan Cheng, Shiqing Ma, Pin-Yu Chen, Tsung-Yi Ho, et al. Elijah: Eliminating backdoors injected in diffusion models via distribution shift. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10847–10855, 2024. 3
- [3] Hiroki Azuma and Yusuke Matsui. Defense-prefix for preventing typographic attacks on clip. *ICCV Workshop on Adversarial Robustness In the Real World*, 2023. 2, 3
- [4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1
- [5] Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *arXiv preprint arXiv:2403.04279*, 2024. 1
- [6] Weixin Chen, Dawn Song, and Bo Li. Trojdiff: Trojan attacks on diffusion models with diverse targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4035–4044, 2023. 3
- [7] Hao Cheng, Jinhao Duan, Hui Li, Lyutianyang Zhang, Jiahang Cao, Ping Wang, Jize Zhang, Kaidi Xu, and Renjing Xu. Rbformer: improve adversarial robustness of transformer by robust bias. *The British Machine Vision Conference (BMVC)*, 2023. 3
- [8] Hao Cheng, Erjia Xiao, Jindong Gu, Le Yang, Jinhao Duan, Jize Zhang, Jiahang Cao, Kaidi Xu, and Renjing Xu. Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language model. *European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 8
- [9] Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villandiffusion: A unified backdoor attack framework for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 3
- [10] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8795–8805, 2024. 1
- [11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 1, 3
- [12] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3
- [13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 5
- [14] Sensen Gao, Xiaojun Jia, Yihao Huang, Ranjie Duan, Jindong Gu, Yang Liu, and Qing Guo. Rt-attack: Jailbreaking text-to-image models via random token. *arXiv preprint arXiv:2408.13896*, 2024. 3
- [15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2
- [16] Jindong Gu. Responsible generative ai: What to generate and what not. *arXiv preprint arXiv:2404.05783*, 2024. 3
- [17] Laura Hanu and Unitary team. Detoxify. Github. <https://github.com/unitaryai/detoxify>, 2020. 2, 3, 7
- [18] Cheng Hao, Kaidi Xu, Sijia Liu, Pin-Yu Chen, Pu Zhao, and Xue Lin. Defending against backdoor attack on deep neural networks. *arXiv preprint arXiv:2002.12162*, 2020. 3
- [19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3, 5
- [20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 3
- [21] Seunghoo Hong, Juhun Lee, and Simon S Woo. All but one: Surgical concept erasing with model preservation in text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21143–21151, 2024. 3
- [22] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*, 2023. 1, 7
- [23] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 3, 4, 5
- [24] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 8, 1
- [25] Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12006–12016, 2024. 3
- [26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 3
- [27] Yunchen Li, Zhou Yu, Gaoqi He, Yunhang Shen, Ke Li, Xing Sun, and Shaohui Lin. Spd-ddpm: Denoising diffusion probabilistic models in the symmetric positive definite

- space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13709–13717, 2024. 3
- [28] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. *arXiv preprint arXiv:2302.04578*, 2023. 3
- [29] Fengyuan Liu, Haochen Luo, Yiming Li, Philip Torr, and Jindong Gu. Which model generated this image? a model-agnostic approach for origin attribution. In *European Conference on Computer Vision*, pages 282–301. Springer, 2024. 2
- [30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 3
- [31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 1, 3
- [32] Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. Latent guard: a safety framework for text-to-image generation. In *European Conference on Computer Vision*, pages 93–109. Springer, 2025. 2, 3, 7
- [33] Tong Liu, Zhixin Lai, Gengyuan Zhang, Philip Torr, Vera Demberg, Volker Tresp, and Jindong Gu. Multimodal pragmatic jailbreak on text-to-image models. *European Conference on Computer Vision (ECCV)*, 2024. 3
- [34] Xin Liu, Yichen Zhu, Yunshi Lan, Chao Yang, and Yu Qiao. Query-relevant images jailbreak large multi-modal models. *arXiv preprint arXiv:2311.17600*, 2023. 3
- [35] Ziwei Liu, Ping Luo, Xiaoqiang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018. 4, 5
- [36] Romain Lopez, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Information constraints on auto-encoding variational bayes. *Advances in neural information processing systems*, 31, 2018. 3
- [37] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *The Twelfth International Conference on Learning Representations*. 3
- [38] Jiachen Ma, Anda Cao, Zhiqing Xiao, Yijiang Li, Jie Zhang, Chao Ye, and Junbo Zhao. Jailbreaking prompt attack: A controllable adversarial attack against diffusion models. *arXiv preprint arXiv:2404.02928*, 2024. 3
- [39] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15009–15018, 2023. 1, 7
- [40] Nithin Gopalakrishnan Nair and Vishal M Patel. Dreamguider: Improved training free diffusion-based conditional generation. *arXiv preprint arXiv:2406.02549*, 2024. 1
- [41] Nithin Gopalakrishnan Nair, Kangfu Mei, and Vishal M Patel. At-ddpm: Restoring faces degraded by atmospheric turbulence using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3434–3443, 2023. 3
- [42] Hyelin Nam, Gihyun Kwon, Geon Yeong Park, and Jong Chul Ye. Contrastive denoising score for text-guided latent diffusion image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9192–9201, 2024. 1
- [43] Yan Pang, Tianhao Wang, Xuhui Kang, Mengdi Huai, and Yang Zhang. White-box membership inference attacks against diffusion models. *arXiv preprint arXiv:2308.06405*, 2023. 3
- [44] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5
- [45] Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara, et al. Safe-clip: Removing nsfw concepts from vision-and-language models. In *Proceedings of the European Conference on Computer Vision*, 2024. 2, 3
- [46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3
- [47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5
- [48] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3, 5
- [49] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022. 2, 3, 7, 8
- [50] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 3
- [51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 5
- [52] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023. 3
- [53] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015. 3

- [54] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *The Twelfth International Conference on Learning Representations*, 2023. 1, 3
- [55] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024. 8, 1
- [56] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1
- [57] Yihan Wen, Xianping Ma, Xiaokang Zhang, and Man-On Pun. Gcd-ddpm: A generative change detection model based on difference-feature guided ddpm. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3
- [58] Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. *arXiv preprint arXiv:2412.13817*, 2024. 3
- [59] Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE symposium on security and privacy (SP)*, pages 897–912. IEEE, 2024. 3
- [60] Hu Ye, Jun Zhang, Sibao Liu, Xiao Han, and Wei Yang. Ip-adapt: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 1, 3, 5
- [61] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10146–10156, 2023. 1
- [62] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Conference on Neural Information Processing Systems*, 2024. 3
- [63] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 3