CLIPAG: Towards Generator-Free Text-to-Image Generation

Roy Ganz
Department of ECE
Technion, Haifa, Israel
ganz@campus.technion.ac.il

Michael Elad
Department of Computer Science
Technion, Haifa, Israel
elad@cs.technion.ac.il

Abstract

Perceptually Aligned Gradients (PAG) refer to an intriguing property observed in robust image classification models, wherein their input gradients align with human perception and pose semantic meanings. While this phenomenon has gained significant research attention, it was solely studied in the context of unimodal vision-only architectures. In this work, we extend the study of PAG to Vision-Language architectures, which form the foundations for diverse image-text tasks and applications. Through an adversarial robustification finetuning of CLIP, we demonstrate that robust Vision-Language models exhibit PAG in contrast to their vanilla counterparts. This work reveals the merits of CLIP with PAG (CLIPAG) in several vision-language generative tasks. Notably, we show that seamlessly integrating CLIPAG in a "plug-n-play" manner leads to substantial improvements in vision-language generative applications. Furthermore, leveraging its PAG property, CLIPAG enables text-to-image generation without any generative model, which typically requires huge generators.

Figure 1. **Unimodal Perceptually Aligned Gradients**. Visualizations of large-ϵ targeted adversarial attacks on non-robust and robust ResNet-50, trained on ImageNet. Such attacks lead to semantically meaningful modifications in the robust model, indicating the generative capabilities of models with PAG. Contrary, the modifications done by the "vanilla" one are entirely meaningless.

1. Introduction

Adversarial robustness is an essential objective in deep learning, requiring models to be insensitive to small malicious input perturbations, referred to as adversarial attacks. Tsipiras et al. [59] discovered a surprising property of adversarially robust models, commonly referred to as Perceptually Aligned Gradients (PAG). According to this trait, the input gradients of the model with respect to a specific class are semantically related to it, being significantly more aligned to human perception than non-robust ones. An implication of this is that models with PAG have generative capabilities that can be leveraged using pixel space optimization. Specifically, the outputs of strong targeted adversarial attacks lead to modifications that perceptually correlate with the target class (see Figure 1). This exciting phenomenon has gained much research attention, with attempts at better understanding it [17, 24, 33, 57] and harnessing it for various computer vision applications [5, 23, 53]. Interestingly, PAG has been explored so far solely in the context of unimodal vision-only applications.

In contrast to the existing PAG literature that primarily focuses on unimodal vision-only applications, our work delves into the domain of Vision-Language (VL) models, which is gaining significant research and attention these days [25, 36, 37, 44, 50, 61, 62, 67]. Our exploration focuses on CLIP [44] – Contrastive Language-Image Pretraining – a powerful Vision-Language model that learns a joint feature space for images and their captions. Building upon the understandings from the unimodal PAG research [11, 33, 59], we consider an adversarial finetuning of the visual part of CLIP as a method that can potentially induce gradient alignment. We demonstrate that while "vanilla" CLIP does not possess PAG at all, its robust counterpart does (see Figure 3). We denote the resulting model as **CLIPAG** – **CLIP** with **P**erceptually **A**ligned **G**radients, and show experimentally that adversarial training in this VL model implies PAG, as in unimodal ones.

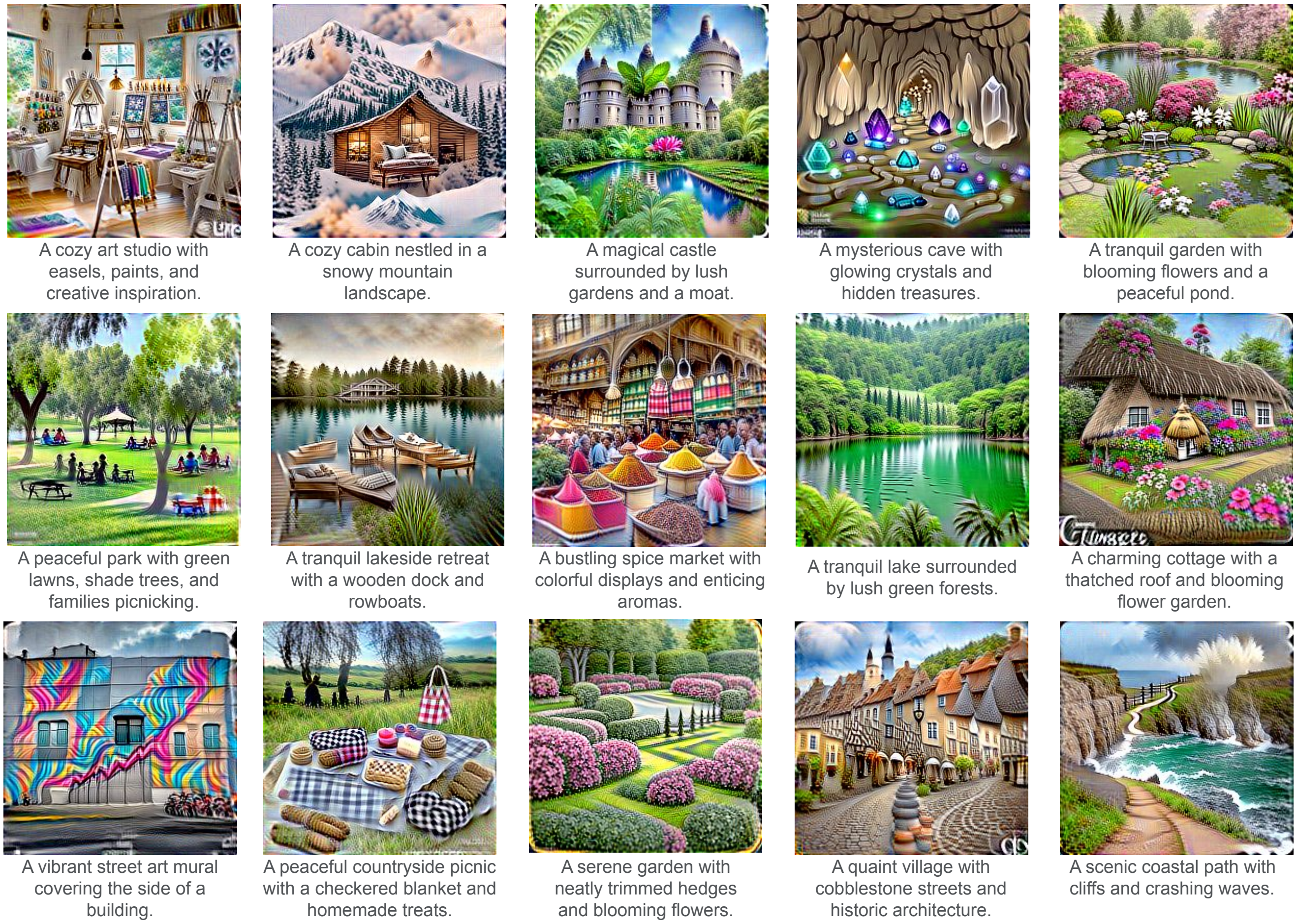

Figure 2. **CLIPAG generator-free text-to-image generation.**

Ever since its introduction, CLIP has become a foundational model for a wide range of text-to-image generative tasks [12, 20, 22, 31, 34, 41, 43, 45, 60]. These often involve modifying images to maximize the alignment with a given text prompt, achieved by deriving CLIP's vision encoder and updating the image to maximize the cosine similarity with the text. However, CLIP is known to be vulnerable to adversarial attacks [19] and lacks PAG, which poses a significant challenge in achieving the desired meaningful visual modifications. Indeed, CLIPDraw [20], a CLIP-based text-to-drawing framework, acknowledged this limitation, stating "*synthesis through-optimization methods often result in adversarial images that fulfill the numerical objective but are unrecognizable to humans*". To mitigate this, researchers have developed ad-hoc techniques and tricks to regularize and improve CLIP gradients, such as optimizing a generator's latent space [12, 22, 34, 43] or utilizing Bézier curves rather than operating in the pixel-domain [20, 60]. In this context, our proposed CLIPAG is a natural solution to this limitation, as discussed next.

We embark on demonstrating the benefits of CLIPAG by seamlessly integrating it into existing CLIP-based generative frameworks in a "plug-n-play" manner. Specifically, we consider both text-to-image generative tasks using CLIPDraw [20] and VQGAN+CLIP [12] and text-based stylization using CLIPStyler [34]. We show that replacing CLIP with CLIPAG leads to improved performance in all these fronts. Interestingly, CLIPAG alleviates the need for gradient regularization techniques, offering a more straightforward approach for leveraging CLIP in such tasks.

Inspired by the above, we propose a novel text-to-image generation via a simple iterative framework using CLIPAG. Unlike the above-described experiments in which CLIPAG is merged into existing solutions, this synthesis framework is a direct *pixel-domain-based* approach. Amazingly, and in contrast to existing text-to-image methods that rely on huge generative networks (*e.g.*, diffusion and GANs), CLIPAG enables high-quality pixel-space image generation (see examples in Figure 2) without any explicit training of a generative model and while using a small pretrained network.

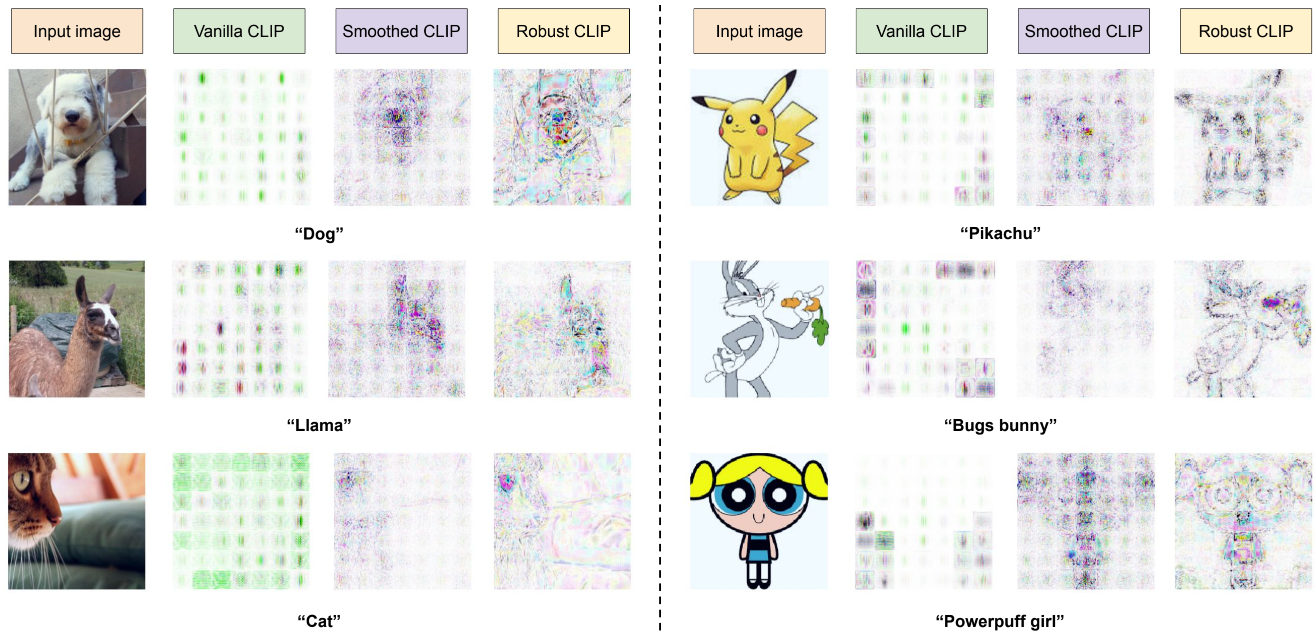

Figure 3. **PAG in CLIP.** A demonstration of the PAG phenomenon using CLIP. We present the input-gradients w.r.t. a given text of both the “vanilla” CLIP, randomized-smoothed CLIP, and CLIPAG in a zero-shot setting for natural images from ImageNet [13] (left) and cartoons (right). While the former's gradients are completely meaningless, the latters lead to gradients that better align perceptually with the given text. Specifically, the adversarially robust CLIP leads to better alignment.

To summarize, the main contributions of this work are as follows: (i) We introduce CLIPAG - an adversarially fine-tuned version of the visual encoder of CLIP that exhibits perceptually aligned gradients; (ii) We integrate CLIPAG into existing text-to-image frameworks, substantially improving their performance while also simplifying these algorithms; and (iii) We leverage CLIPAG to propose a simple, generator-free text-to-image synthesis solution, producing high-quality synthesized images.

2. Related Work

2.1. Adversarial Attacks and Robustness

Adversarial Attacks Given an image classifier $f_\theta(\mathbf{x})$, adversarial examples are crafted by attackers in order to fool it and divert the classification decision. It was discovered that adding a small imperceptible noise δ to an image can lead to misclassification [26, 58]. In practice, the adversarial noise is constrained within a defined threat model Δ, often defined as a small norm ball (*e.g.*, $\Delta = \{\delta \ : \ \|\delta\|_\infty \leq \frac{8}{255}\}$). Formally, given an input sample $\mathbf{x}$, its true label y, and a threat model Δ, a valid adversarial example $\hat{\mathbf{x}}$ satisfies the following conditions: $\hat{\mathbf{x}} = \mathbf{x} + \delta \ s.t. \ \delta \in \Delta$ and $y_{\text{pred}} \neq y$, where y_{pred} is the predicted label by the classifier f_θ for $\hat{\mathbf{x}}$. The process of generating such examples is called adversarial attacks, and numerous methods have been developed for this purpose [7, 15, 26, 39]. This work focuses on the Projected Gradient Descent (PGD) method [39].

Adversarial Robustness The above-described vulnerability of classifiers sparked research dedicated to enhancing their robustness against such attacks. A commonly considered solution is adversarial training [26, 39], which approximates the solution of the following min-max optimization:

$$\min_\theta \sum_{(\mathbf{x},y)\in\mathcal{D}} \max_{\delta\in\Delta} \mathcal{L}(f_\theta(\mathbf{x}+\delta), y), \tag{1}$$

in which the classifier is trained to correctly classify the most challenging adversarial examples allowed by the threat model Δ. An additional effective technique for robustifying neural networks is randomized smoothing [11], in which the classifier is smoothed by convolution with Gaussian noise. Specifically, in the L_2 case, the robust classifier $\hat{f}_{\theta,\sigma}$ is a smoothed version of f_θ,

$$\hat{f}_{\theta,\sigma} = \mathbb{E}_{\mathbf{n}\sim\mathcal{N}(0,\sigma^2\mathbf{I})}[f_\theta(\mathbf{x}+\mathbf{n})], \tag{2}$$

where σ controls the robustness-accuracy tradeoff.

2.2. Perceptually Aligned Gradients

Perceptually aligned gradients (PAG) [17, 18, 59] refer to classifier input-gradients, $\nabla_\mathbf{x} f_\theta(y|\mathbf{x})$, that are semantically aligned with human perception. Consequently, when an image is altered to maximize the probability of a specific class in a model with PAG, the modifications made to the image are semantically meaningful, as demonstrated in Figure 1. PAG has been found to exist in adversarially trained models

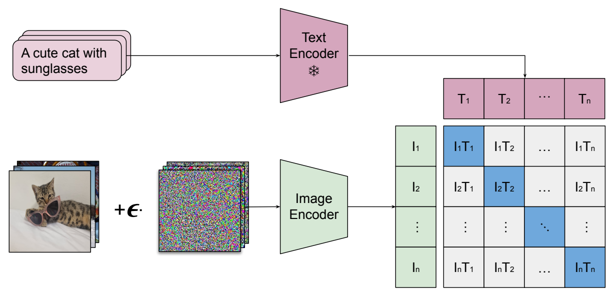

Figure 4. **Vision-Language Contrastive Adversarial Training.** According to Equation 3, we first craft adversarial perturbation δ to minimize the similarity between the corresponding image-text pairs, *i.e.*, reduce the values on the main diagonal of the depicted matrix. Next, we train the CLIP model to align the adversarial examples and their matching text, *i.e.*, to maximize the values on the main diagonal. Repeating these steps results in adversarially robust CLIP that possesses Perceptually Aligned Gradients.

but not in "vanilla" ones [59], indicating that the features learned by the former are more aligned with human vision.

PAG has recently drawn significant research attention. Theoretically oriented studies have focused on better understanding this trait and the circumstances for its appearance. By demonstrating PAG in randomized smoothed classifiers, it has been established that PAG is a general property of robust models and not solely an artifact of adversarial training [33]. Furthermore, Ganz *et al.* [24] demonstrated the bidirectional relationship between PAG and robustness, revealing that PAG implies robustness and vice-versa. Additionally, Srinivas *et al.* [57] recently shed some light on the cause of PAG via off-manifold robustness analysis. Applicative-oriented studies aimed at leveraging PAG for generative tasks, such as image generation and image-to-image translation [53], improving state-of-the-art image generation results [23]. PAG has also been explored for improved robust classification [5].

Despite this extensive line of work and the growing interest in multimodal networks among the computer vision research community, PAG has only been studied within the context of unimodal vision-only architectures and has never been explored in the vision-language domain. In this work, we aim to close this gap and investigate the existence of PAG in Vision-Language multimodal architectures and its connection to adversarial robustness using both adversarial training and randomized smoothing. Armed with this, we explore the potential of multimodal PAG in improving multiple text-to-image generative applications.

2.3. CLIP in Vision-Language Generative Tasks

CLIP (Contrastive Language-Image Pretraining) [44] is a multimodal Vision-Language model pretrained to align a massive corpus of 400 million pairs of images and their captions. The outstanding richness of CLIP's visual and textual space has been leveraged for various text-to-image generative tasks. Several studies have focused on image generation conditioned on textual descriptions, mainly by guiding the visual result to be aligned with the given text in CLIP's space [12, 20, 31, 41, 45, 48]. VQGAN+CLIP [12] proposed a training-free method to generate images from text by optimizing the latent code of a pretrained VQGAN [65] to output an image that matches the textual description in the CLIP space. Similarly, Ponce *et al.* introduced Clip-DRAW [20], a method that combined CLIP and Bézier curves for generating drawings from texts by optimizing the parameters of such curves for best alignment. Interestingly, [20] has identified CLIP's susceptibility to adversarial attacks: "*A key issue in synthesis through optimization is that the produced images often leave the space of natural images, or fool the system through adversarial means*". Thus, such works mitigated this by avoiding pixel-domain optimization and performing multiview augmentations.

Another line of work involving CLIP is text-guided style transfer and image editing. StyleCLIP [43] and StyleGAN-NADA [22] proposed to leverage a pretrained StyleGAN [32] model with CLIP to adjust the style of images to match a given textual descriptions. CLIPStyler [34] tackles a similar task using a framework that consists of several CLIP-based losses, augmentation pipes, and a style network being trained for each image. CLIP was also used for localized text-based image-editing using internal learning [3]. These studies demonstrate the broad applicability of CLIP in text-to-image generative tasks.

3. Obtaining Vision-Language PAG

In this paper, we delve into the concept of Perceptually Aligned Gradients (PAG) within vision-language models. We focus on the image encoder's input gradients with respect to a given textual input, aiming for structured content that is semantically correlated with the text. To this end, we leverage the well-established observation that both adversarial training and randomized smoothing lead to aligned gradients in unimodal vision-only models [33,59]. First, we explore the gradients of "vanilla" CLIP, using the publicly available CLIP ViT-B/32 [29], using both natural images from ImageNet [13] and arbitrary cartoon ones. To this end, we obtain input gradients by deriving the image encoder to maximize the cosine similarity with a given text in CLIP's feature space. As can be seen in Figure 3, the standard CLIP model has no alignment with the semantically meaningful features while also exhibiting a strong blockiness effect due to the Vision Transformer (ViT) architecture [16].

Next, we examine the gradient alignment of the same CLIP model with randomized smoothing. This approach mitigates the blockiness effect and improves the alignment, as seen in Figure 3, but only to some extent. Hence, we propose to adversarially finetune the CLIP model to improve

Figure 5. **CLIPDraw results.** Visualization of CLIPDraw outputs with CLIP and CLIPAG using different textual prompts styles (artistic in blue, abstract concepts in purple, and realistic in green. The top two rows represent results without augmentation. As can be seen, in this case, CLIP completely fails to guide the optimization process toward meaningful outputs. However, CLIPAG significantly outperforms it, leading to improved drawings that align with the textual description. Moreover, when applying augmentations (bottom rows), CLIPAG still leads to better visual outputs, as also indicated in the quantitative evaluation presented in Table 1.

the PAG property. To this end, we adopt adversarial training techniques [39], as illustrated in Figure 4. We denote the Text Encoder and Image Encoder as $f^T_{\theta_T}$ and $f^I_{\theta_I}$, respectively, where θ_T and θ_I represent the models' parameters. Given an image $\mathbf{x}$ and its corresponding caption t from an image-text dataset $\mathcal{D}$, we first craft adversarial example $\mathbf{x}+\delta$ aimed to minimize the similarity between matching image-caption pairs. Subsequently, we update the model weights to maximize the similarity between the adversarial examples and their corresponding captions. Formally, we propose solving the following optimization problem, which extends adversarial training to the multimodal case:

$$\min_{\theta_I,\theta_T} \sum_{(\mathbf{x},t)\in\mathcal{D}} \max_{\delta\in\Delta} \mathcal{L}_{SIM}(f^I_{\theta_I}(\mathbf{x}+\delta), f^T_{\theta_T}(t)), \quad (3)$$

where $\mathcal{L}_{SIM}$ represents cosine similarity loss calculated in CLIP's feature space. We visualize the optimization process described in Equation 3 in Figure 4 while taking into account the fact that CLIP is trained over batches of pairs via contrastive learning.

Conducting such training brings several challenges and design choices. First, CLIP was originally trained on massive 400 million image-text pairs with a humongous batch size of $32,768$, using hundreds of GPUs. The combination of the training set size, which introduces a huge diversity, and the large mini-batches contributing to the effectiveness of contrastive training led to unprecedented capabilities. Finetuning such an architecture using academic resources might potentially lead to catastrophic forgetting [21] and deteriorate performance. Moreover, when finetuning on adversarial examples, this could be further exacerbated. The vulnerability of CLIP to adversarial attacks might force a massive change of parameters during the finetuning, resulting in significant degradation of the generalization capabilities of CLIP, which are necessary for a wide range of downstream generative tasks. However, recall that our objective is not to robustify CLIP but rather to align its gradients. Considering the challenges mentioned above, and the observation that in the unimodal case, adversarial training with even a low maximum perturbation bound can lead to perceptually aligned gradients [2], we focus on adversarial training using a small threat model. We hypothesize that this strategy will lead to PAG with less detrimental effects than applying adversarial training with the common robustification threat models. In practice, we use a threat model of $\Delta = \{\delta : \|\delta\|_2 \leq 1.5\}$ for images of $224\times 224\times 3$, which is approximately equivalent to a mean pixel change of $\frac{1}{255}$, which is much smaller than the ones used in adversarial robustness research.

In our approach, we focus on aligning the gradients of the image encoder, and thus, we choose to freeze the parameters of the pretrained text encoder when solving Equation

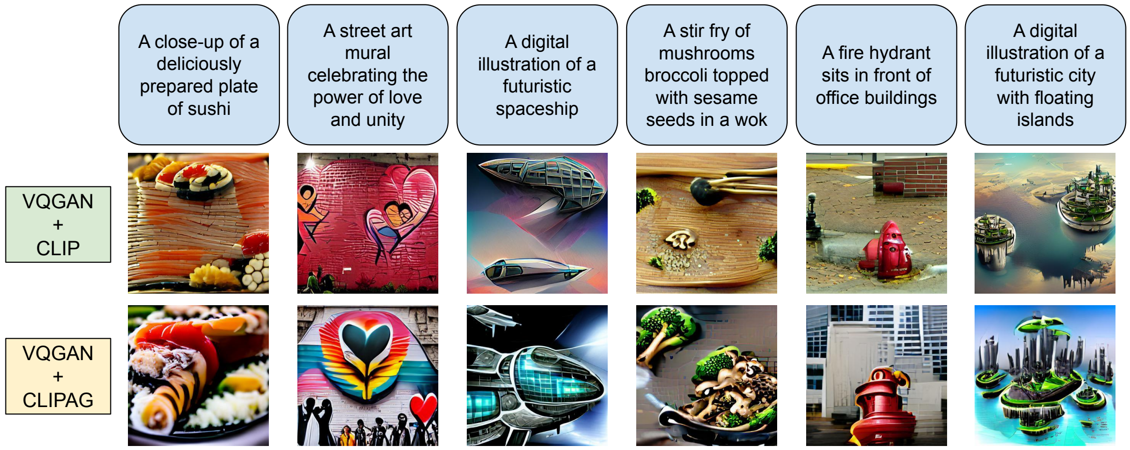

Figure 6. **VQGAN+CLIP results.** Visualization of VQAGAN + CLIP results using both CLIP and CLIPAG on various text prompts using the same hyperparameters and seed. As can be seen, seamlessly integrating CLIPAG to VQGAN+CLIP improves the generated images.

3. By doing so, we not only reduce computational costs (freezing half of the model's parameters) but also introduce a valuable stabilizing mechanism during the adversarial finetuning of the image encoder. This is particularly effective because the pretrained text encoder generates meaningful representations, and aligning the image encoder with it can significantly enhance its performance. Furthermore, we conduct comprehensive experiments to thoroughly evaluate the impact of various design choices, which we discuss in the supplementary materials. Specifically, we investigate the effects of different architectures, including both convolutional networks and Vision Transformers (ViT) [16], as well as different threat models, such as L_2 and L_∞ norms. In our practical implementation, we mainly focus on CLIP's ViT-B/32 architecture, which is widely utilized in downstream tasks and thus enables a fair comparison. We train it using a concatenated dataset that combines SBU [42], CC3M [56], CC12M [8] and LAION-400M [54] and we subsample them to obtain a uniform dataset (*i.e.*, the distribution to draw from each source is approximately the same). We perform a short finetuning with a low learning rate for the model using eight A40 GPUs while keeping the text encoder frozen (implementation details listed in Appendix A).

After adversarially finetuning CLIP, we assess whether it achieves greater gradient alignment than the "vanilla" model and the randomized smoothed one. As illustrated in Figure 3, the robustly trained variant exhibits the highest alignment with the provided texts. These findings indicate that similar to the unimodal case, robustification yields PAG in the context of VL models. Due to its superiority, we focus hereafter on the adversarially trained model.

4. CLIPAG as a Generative Model

In this section, we explore the benefits of CLIPAG (CLIP with Perceptually Aligned Gradients) in various generative tasks in two main settings – CLIP-based generative frameworks and generator-free text-to-image generation.

4.1. Text-to-Image Generative Frameworks

Due to its strong vision-language alignment, CLIP is used as a fundamental block in various text-to-image generative tasks and applications, such as text-based editing [4, 22, 34, 43, 60] and generation [12, 20, 31, 41, 45, 48]. In this section, we demonstrate that CLIPAG can be integrated into existing text-to-image generation applications in a "plug-n-play" manner by simply replacing the "vanilla" CLIP with its robust counterpart. Specifically, to thoroughly explore the effects of leveraging CLIPAG, we experiment with it in text-based image editing and generation frameworks by integrating it into CLIPDraw [20], VQGAN+CLIP [12] and CLIPStyler [34]. We focus mainly on CLIPDraw due to its simplicity and lack of a generative model, which enables us to explore the generative capabilities of CLIPAG compared to the standard CLIP. In addition, in Appendix B, we demonstrate that besides generative tasks, CLIPAG can lead to improved explainability.

Text-based Image Generation In this setting, we consider both CLIPDraw [20] and VQGAN+CLIP [12]. CLIPDraw proposed a CLIP-based approach for text-to-drawing generation by optimizing a set of Bézier curves to minimize the cosine distance in the CLIP space between generated images and description prompts. Additionally, to overcome the issue of the non-aligned gradients of CLIP, the authors utilized a multiview augmentation pipeline. As stated in their paper, "*without image augmentation, synthesis through-optimization methods often result in adversarial images that fulfill the numerical objective but are unrecognizable to humans*." We study the effect of replacing CLIP in CLIPDraw with CLIPAG in two configurations – with

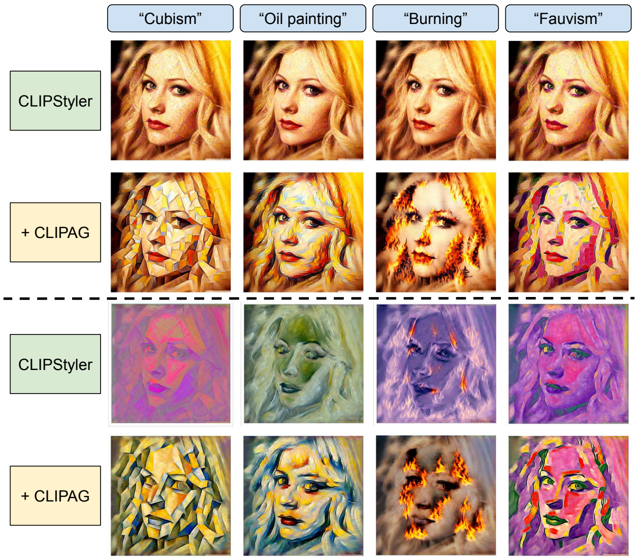

Figure 7. **CLIPStyler results.** Visualization of applying CLIPStyler [34] using both CLIP and CLIPAG, with and without the style network (bottom and top, respectively). As can be seen, while CLIP struggles to guide the style transfer process without a style network, CLIPAG leads to much better results, attesting to its improved gradients. Additionally, even with the introduction of the style network, the outputs of CLIPAG are more convincing.

Method	Aug.	Aesthetics		Caption consistency			
				ConvNext		ViT-H/14	
		Score	Pref.	Sim.	R-Prec	Sim.	R-Prec
CLIPDraw [20]	✗	3.70	22%	27.1	2%	18.1	1%
+ CLIPAG	✗	**3.98**	**78%**	**32.1**	**19%**	**26.3**	**32%**
CLIPDraw [20]	✓	4.14	34%	35.4	31%	31.6	70%
+ CLIPAG	✓	**4.31**	**66%**	**36.0**	**52%**	**32.0**	**72%**

Table 1. **CLIPDraw quantitative results.** CLIPDraw results using CLIP and CLIPAG using aesthetic and caption consistency metrics, with and without augmentation pipeline. As can be seen, simply replacing CLIP with CLIPAG leads to a substantial improvement in terms of aesthetics and caption similarity, attesting to the benefits of CLIP with Perceptually Aligned Gradients.

and without the augmentation pipe. The rationale behind omitting the augmentation is to compare the generative capabilities of CLIP and CLIPAG. First, we analyze the performance qualitatively and present the results in Figure 5. As can be seen, CLIPAG leads to improved performance compared to the baseline and is also capable of operating without augmentation, unlike the baseline, due to PAG.

In order to quantitatively analyze performance, we adopt an automatic procedure for aesthetics assessment. To this end, we generate 100 artistic prompts using ChatGPT [6] by conditioning on prompt examples from CLIPDraw's paper as a prefix. Next, we generate 100 images using CLIPDraw with the baseline and CLIPAG. Finally, we propose two main empirical metrics for assessing the performance – (i) Utilize a linear-probed CLIP model on AVA dataset [40], a human-annotated dataset of aesthetics containing over $250,000$ images (the aesthetic scores are between 1 to 10). A similar technique was also adopted in [45]. Based on this model, we report the mean aesthetic score and the Preference rate between CLIPDraw using CLIP and CLIPAG. (ii) We utilize two publicly-available CLIP models [29] (ConvNext[1] and ViT-H/14[2]) and report two metrics aimed at capturing the caption consistency – the cosine similarity (Sim.) and the retrieval precision (R-Prec). R-Prec is averaged precision of a retrieval task where, for each generated image, the model predicts the most probable prompt among all the 100 generated prompts. We report the above scores in Table 1 both with and without augmentation. As can be seen, CLIPAG significantly outperforms the baseline in all metrics when omitting the augmentations due to the PAG property. Interestingly, even with the augmentations aimed at mitigating CLIP's susceptibility to adversarial attacks, CLIPAG leads to much-improved performance.

An additional evidence of the benefits of CLIPAG in text-based generative frameworks is provided by using VQGAN+CLIP [12]. In Figure 6, we qualitatively demonstrate the effectiveness of replacing CLIP with CLIPAG in the VQGAN+CLIP framework on different prompts. To quantitatively evaluate the effect of CLIPAG in the VQGAN+CLIP framework, we randomly sample 100 captions from the validation set of MS-COCO captions [10], generate images accordingly with both CLIP and CLIPAG and calculate their cosine similarity using OpenCLIP ViT-H/14. Despite the fact that OpenCLIP is not robust and probably more aligned with the "vanilla" one, CLIPAG obtains **34.3** cosine similarity, surpassing CLIP's **33.4**. We provide additional details regarding these experiments in Appendix A, along with additional qualitative results.

Text-based Image Editing For image editing tasks, we adopt CLIPStyler [34] as our framework, which leverages a pretrained CLIP model for text-based image style transfer. CLIPStyler addresses the issue of meaningless CLIP gradients by incorporating a style network and employing a multiview augmentation pipeline. In this study, we compare the performance of CLIPStyler using the standard CLIP and CLIPAG model in two settings: (i) the "plug-n-play" approach and (ii) without the style network (to isolate the guidance capability of CLIPAG and CLIP in style-transfer). As can be seen in Figure 7, seamlessly replacing CLIP with CLIPAG leads to marked improvement. In addition, while utilizing CLIP without the style network completely fails, CLIPAG's results are substantially better, showcasing its capabilities in guiding a text-based style transfer process.

[1] CLIP-convnext-base-w-320-laion-aesthetic-s13B-b82K
[2] CLIP-ViT-H-14-laion2B-s32B-b79K

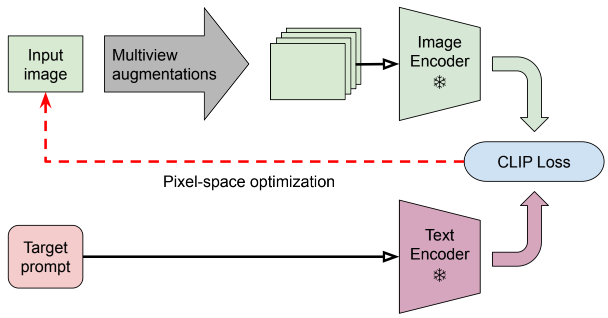

Figure 8. **Generator-free text-to-image framework.** Visualization of the proposed method to harness CLIPAG for text-to-image generation without any generative model. Specifically, we iteratively update the pixels of an image to better align to a textual description. Due to PAG, such a process leads to meaningful images that align with the text.

4.2. Generator-Free Text-to-Image Generation

In recent years, text-to-image generation has gained substantial attention, leading to the development of various methods aimed at tackling this challenging task [12, 35, 45, 47, 49, 51]. While these methods employ diverse techniques, they all rely on powerful generative models. In this section, we present a novel approach that breaks away from traditional text-to-image pipelines by leveraging CLIPAG, a compact, non-generative model with approximately 150 million parameters. To this end, we introduce a simple yet effective *generator-free text-to-image synthesis* framework, as depicted in Figure 8, which employs an iterative pixel-space optimization technique to align an input image with a target text, leveraging the Perceptually Aligned Gradients of CLIPAG. Notably, our approach does not involve training any of CLIPAG's components. The proposed framework consists of three key blocks: (i) **Initialization** – Sampling the input for the generation process. To achieve this, we model a low-resolution dataset as a Gaussian Mixture Model and select the candidate image with the best alignment to the given text based on CLIP; (ii) **Multiview augmentations** – in each iteration, we duplicate the image and perform differential random augmentations. In practice, we use DiffAugment [68] and random cropping; (iii) **CLIPAG-based Loss** – in every step, we update the input image using the input gradients of CLIPAG to better align with the text. We provide the implementation details in Appendix A.

To qualitatively demonstrate the capabilities of our generator-free framework, we visualize generated images corresponding to various textual descriptions in Figure 2 and Appendix D. The images produced by CLIPAG exhibit a high level of visual fidelity and consistency with the given text, despite the absence of a traditional generative model. For quantitative assessment, we conduct experiments on the MS-COCO dataset. First, we synthesize the same 100 prompts used in the VQGAN+CLIP experiments and calculate the CLIPScore using OpenCLIP ViT-14/H. Our framework achieves an impressive score of **36.8**, surpassing the performance of the generator-based framework. Specifically, we outperform VQGAN with CLIP by ↑ **3.4** and VQGAN with CLIPAG by ↑ **2.5**. Next, to further evaluate the performance, we generate 30,000 images from MS-COCO validation captions in a zero-shot (ZS) setting. We calculate the CLIPScore [27], Inception Score (IS) [52], and Frechet Inception Distance (FID) [28], and compare our results against strong baselines in Table 2. Notably, while using a significantly smaller non-generative model, our approach outperforms DALL-E [47] and CogView [14] in Inception Score. However, our FID metric is relatively weaker, potentially attributed to CLIPAG's tendency to generate colorful and artistic images, which deviate from the characteristics of the MS-COCO. We hypothesize that this can be mitigated via prompt tuning but leave this for future work. In addition, we explore different aspects of our proposed scheme, in Appendix D. While our results do not yet rival state-of-the-art methods [41, 46, 51], they highlight the remarkable generative capabilities of CLIPAG and potentially pave the way for a new family of generator-free text-to-image generation techniques.

Method	#Params.	ZS	IS↑	FID↓	CLIPScore↑
Stack-GAN [66]	-	✗	8.5	74.1	-
AttnGAN [64]	230M	✗	**23.3**	35.5	27.7
CogView [14]	4,000M	✗	18.2	<u>27.1</u>	<u>33.2</u>
DALL-E [47]	12,000M	✓	17.9	27.5	-
GLIDE [41]	6,000M	✓	-	**12.2**	-
Ours	150M	✓	<u>18.7</u>	42.3	**34.7**

Table 2. **MS-COCO text-to-image generation results**. Inception score, CLIPScore (higher is better), and Frechet Inception Distance (lower is better) results, along with model sizes.

5. Discussion and Conclusions

In this paper, we explore the concept of PAG in the context of Vision-Language architectures using CLIP. Our findings highlight several significant contributions. First, we establish the presence of PAG in CLIP by adversarially fine-tuning it. This demonstrates that the phenomenon of PAG is not limited to unimodal vision-only architectures but extends to multimodal models. Second, we demonstrate that CLIPAG can be seamlessly integrated into existing text-to-image existing frameworks, leading to substantial improvements. Lastly, we showcase that CLIPAG can be used for generator-free text-to-image synthesis, which typically relies heavily on generative models. Our results demonstrate the practical implications and potential of harnessing PAG in real-world Vision-Language applications. We believe the insights and findings presented in this paper will inspire further exploration and advancements in harnessing PAG in multimodal research.

References

[1] Aviad Aberdam, Roy Ganz, Shai Mazor, and Ron Litman. Multimodal semi-supervised learning for text recognition. *CoRR*, abs/2205.03873, 2022. 15

[2] Gunjan Aggarwal, Abhishek Sinha, Nupur Kumari, and Mayank Singh. On the benefits of models with perceptually-aligned gradients, 2020. 5

[3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing, 2022. 4

[4] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XV*, volume 13675 of *Lecture Notes in Computer Science*, pages 707–723. Springer, 2022. 6, 13

[5] Tsachi Blau, Roy Ganz, Chaim Baskin, Michael Elad, and Alexander M. Bronstein. Classifier robustness enhancement via test-time transformation. *CoRR*, abs/2303.15409, 2023. 1, 4

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are unsupervised multitask learners. *arXiv preprint arXiv:2005.14165*, 2020. 7, 12

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*, 2017. 3

[8] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021. 6

[9] Peijie Chen, Qi Li, Saad Biaz, Trung Bui, and Anh Nguyen. gscorecam: What objects is clip looking at? In *Proceedings of the Asian Conference on Computer Vision*, pages 1959–1975, 2022. 13

[10] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 7

[11] Jeremy M Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing, 2019. 1, 3

[12] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. VQGAN-CLIP: open domain image generation and editing with natural language guidance. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVII*, volume 13697 of *Lecture Notes in Computer Science*, pages 88–105. Springer, 2022. 2, 4, 6, 7, 8

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 4

[14] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers, 2021. 8

[15] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Conference on Computer Vision and Pattern Recognition, CVPR*, 2018. 3

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. 4, 6

[17] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv*, 2019. 1, 3

[18] Christian Etmann, Sebastian Lunz, Peter Maass, and Carola Schoenlieb. On the connection between adversarial robustness and saliency map interpretability. In *International Conference on Machine Learning, ICML*, 2019. 3

[19] Stanislav Fort. Adversarial examples for the openai clip in its zero-shot classification regime and their semantic generalization, Jan 2021. 2

[20] Kevin Frans, Lisa B. Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. In *NeurIPS*, 2022. 2, 4, 6, 7

[21] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. 5

[22] Rinon Gal, Or Patashnik, Haggai Maron, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Trans. Graph.*, 41(4):141:1–141:13, 2022. 2, 4, 6, 13

[23] Roy Ganz and Michael Elad. BIGRoc: Boosting image generation via a robust classifier. *Transactions on Machine Learning Research*, 2022. 1, 4

[24] Roy Ganz, Bahjat Kawar, and Michael Elad. Do perceptually aligned gradients imply robustness? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10628–10648. PMLR, 2023. 1, 4

[25] Roy Ganz, Oren Nuriel, Aviad Aberdam, Yair Kittenplon, Shai Mazor, and Ron Litman. Towards models that can see and read. *CoRR*, abs/2301.07389, 2023. 1

[26] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations, ICLR*, 2015. 3

[27] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 8

[28] 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, 2018. 8

[29] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below. 4, 7, 12

[30] Ajay Jain, Amber Xie, and Pieter Abbeel. Vectorfusion: Text-to-svg by abstracting pixel-based diffusion models, 2022. 12

[31] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4, 6

[32] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020. 4

[33] Simran Kaur, Jeremy M. Cohen, and Zachary C. Lipton. Are perceptually-aligned gradients a general property of robust classifiers? *CoRR*, 2019. 1, 4

[34] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18062–18071, June 2022. 2, 4, 6, 7, 13

[35] Dongxu Li, Junnan Li, and Steven C. H. Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing, 2023. 8

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *CoRR*, abs/2301.12597, 2023. 1

[37] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR, 2022. 1

[38] Ron Litman, Oron Anschel, Shahar Tsiper, Roee Litman, Shai Mazor, and R. Manmatha. Scatter: Selective context attentional scene text recognizer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 15

[39] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations, ICLR*, 2018. 3, 5

[40] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE, 2012. 7

[41] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 2022. 2, 4, 6, 8

[42] Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems (NIPS)*, 2011. 6

[43] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 2065–2074. IEEE, 2021. 2, 4, 6, 13

[44] 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, 2021. 1, 4

[45] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. 2, 4, 6, 7, 8

[46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 8

[47] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 8

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 4, 6

[49] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 8

[50] Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. Fusecap: Leveraging large language models to fuse visual data into enriched image captions. *CoRR*, abs/2305.17718, 2023. 1

[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 8

[52] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016. 8

[53] Shibani Santurkar, Dimitris Tsipras, Brandon Tran, Andrew Ilyas, Logan Engstrom, and Aleksander Madry. *Image Synthesis with a Single (Robust) Classifier*. 2019. 1, 4

[54] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021. 6

[55] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society, 2017. 13, 15

[56] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 6

[57] Suraj Srinivas, Sebastian Bordt, and Hima Lakkaraju. Which models have perceptually-aligned gradients? an explanation via off-manifold robustness, 2023. 1, 4

[58] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations, ICLR*, 2014. 3

[59] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness May Be at Odds with Accuracy. In *International Conference on Learning Representations, ICLR*, 2019. 1, 3, 4

[60] Yael Vinker, Ehsan Pajouheshgar, Jessica Y. Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: semantically-aware object sketching. *ACM Trans. Graph.*, 41(4):86:1–86:11, 2022. 2, 6

[61] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 1

[62] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR, 2022. 1

[63] Jiayu Wu, Qixiang Zhang, and Guoxi Xu. Tiny imagenet challenge. *Technical report*, 2017. 12

[64] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 8

[65] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 4

[66] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks, 2018. 8

[67] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5579–5588. Computer Vision Foundation / IEEE, 2021. 1

[68] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *Advances in Neural Information Processing Systems*, 33:7559–7570, 2020. 8