Textual Exclusion - Improving Diffusion Model Generations under Vague Natural Text Prompts

Punnawish (KK) Thuwajit Dev Mehrotra

Jianrui (Harris) Zhang Suyash Raj

Abstract

Text-to-image diffusion models have achieved remarkable image generation ability from user-inputted prompts. While its generation quality is remarkable, large pre-trained generative models particularly struggle with short, vaguely written natural human text prompts due to the dataset it was trained on favoring quantity over quality. Small, high-quality datasets, on the other hand, are prone to overfitting from fine-tuning due to their size, losing diversity. In this work, we present a fine-tuning technique using small datasets (each labeled with a consistent "style") with high text-image correspondence called "textual exclusion", that improves a pre-trained text-to-image diffusion model to perform with vague text prompts. Textual exclusion combines two integral training components – learning the special tokens to describe each style's distribution, then training the denoising U-Net conditioned on these style's descriptions to inherit the high text-image correlation while preventing overfitting on these styles. Stable diffusion fine-tuned with 600 text-image pairs and textual exclusion outperforms its pre-trained counterpart in generating images under vague prompts while retaining its diversity. When conditioned on each learned style, textual exclusion generates results closer to its source distribution than textual inversion. This technique can also be applied to subject-driven generation with remarkable results, especially in out-of-distribution generation.

1. Introduction

1.1. Rationale

Open source Text-to-image diffusion models, like Stable Diffusion [1], have gathered their popularity for their ability to generate both photo-realistic and surreal images from user-inputted prompts, opening the door for community contributions. While its performance is outstanding, we discovered an underlying issue regarding Stable Diffusion's training process, which has potentially led to flaws in its generated response. By using CLIP [2] latent vectors, Stable Diffusion [1] models were trained on data, counted by the millions and even billions, from the internet, namely LAION [3], in image-text pairs for the model to learn the correlation between each query and each image via contrastive learning. The training data, however, lacks quality despite its quantitative advantage because the text prompts were minimally human-supervised and at many times cannot accurately describe the image itself. CLIP [2] is also known to be flawed as adversarial attacks [4] to the input prompt with malicious code that is meaningless to humans can result in Stable Diffusion [1] generating a grossly different image. These two problems are both the underlying reasons for time-to-time less desirable images being generated.

1.2. Objectives and Contributions

In this paper, we aim to improve the text-to-image generation performance of Stable Diffusion [1]by fine-tuning it with higher-quality data with relatively small quantities (200 per style, three styles), namely image-text pairs whose text prompts were written and supervised by humans. To prevent overfitting due to its size, we propose textual exclusion, a fine-tuning technique composed of two integral components – textual inversion [5] that learns the special tokens (i.e. "An image of a dog in the style of $\langle\rangle$") to describe each style's distribution, and fine-tuning the denoising U-Net while still conditioned on these style's descriptions to ensure the resulting U-Net model learns the enhanced text-image correlation while excluding style-related feature, effectively mitigating overfitting. As a result, textual exclusion exhibits several desirable qualities including

- More accurate generation results under vague prompts.
- High similarity of image generation conditioned on styles, as well as style mixing.
- Applications in subject-driven image generation.

2. Literature Review

2.1. Challenges and Limitations of Text-to-Image Generation

Text-to-image generation is a challenging task that involves creating realistic and diverse images from natural language descriptions. This task has many applications in art, education, entertainment, and communication. However, text-to-image generation also faces many difficulties, such as the ambiguity and complexity of natural language, the quality and quantity of the training data, and the evaluation of the generated images.

One of the main components of text-to-image generation is the text encoder, which converts the natural language input into a meaningful representation that can guide the image generation process. A common approach for the text encoder is to use a vision-language model that learns to align images and texts based on their semantic similarity. However, this approach also has some limitations, such as its vulnerability to adversarial attacks [6] that can manipulate its text-to-image similarity scores with meaningless text inputs and its dependence on the quality and diversity of the training data, which may not reflect the natural language expressions of human users.

Another important component of text-to-image generation is the image generator, which synthesizes images from the text representation. A recent advance in this direction is Stable Diffusion [7], a generative model that uses latent diffusion to produce high-resolution and diverse images from text prompts. Stable Diffusion was trained on the LAION-5b dataset [3], which contains more than 5 billion image-text pairs filtered by a vision-language model. Stable Diffusion outperformed previous models on several text-to-image generation benchmarks, such as COCO, CUB, and Oxford Flowers. However, Stable Diffusion also suffers from some drawbacks, such as its reliance on the text prompts for controlling the image synthesis, which may lead to undesired results when the text prompts are vague or inconsistent, and its exposure to the low-quality data from the LAION-5b dataset, which may contain noisy or inaccurate image-text pairs.

2.2. The Roles of Text Prompts

Latent Diffusion Models (LDMs) [1] are a state-of-the-art text-to-image generation model. Through injecting text prompt encoding (via text-image alignment models like CLIP [2]) into a denoising U-Net, LDMs are able to control the synthesis of its image to align with its given description.

Normally, textual control is done through classifier-free guidance [8], which injects text encoding through a second denoising U-Net and weighting said U-Net with a prompt-free one. This exhibits the model's dependence on text prompts in its generation. Prompt-to-prompt [9] revealed the connections from text prompts via cross-attention map, revealing each token from the text prompt spatially affects a region of generation. This proves vague text prompts to be problematic under a prompt-heavy reliance.

Some techniques, like the Self-Attention Guidance (SAG) [10], aim to improve text-to-image generation with the independence of text prompts. SAGs, in particular, use intermediate self-attention maps of models to control the diffusion process. SAGs in diffusion models have demonstrated remarkable progress in image generation quality with less dependence on text prompts [11].

2.3. Limitations on the Text Encoder

The text encoder, CLIP [2], was trained with a focus on developing an encoding that matches texts and images via their semantic similarity. It was trained on 400 million pairs of (image-text) constructed by OpenAI, where they employed 500,000 queries of the most common English words to generate 20,000 image-text pairs per query. The exact method, source, and inclusion criteria of such data are not revealed in the original paper, but one could reasonably assume a considerable amount of web crawling is done with carefully chosen queries.

CLIP itself is also revealed to be flawed, as unrelated text prompts (without grammatical meaning) can alter CLIP's text-to-image similarity [4].

2.4. Limitations on Dataset

The LAION-5b dataset [3] is the latest iteration of the LAION dataset on which stable diffusion [1] is trained. It contains more than 5 billion CLIP [2]-filtered image-text pairs crawled from the internet.

A notable flaw of the dataset is the filtering of image-text pairs with little semantic connections from the dataset [3]. Given the nature of internet descriptions of images, several prompts are sampled from titles, tags, or specific names that do not provide the necessary descriptions in natural human language (i.e. inaccurate, grammatically incorrect). Figure 1 illustrates this issue given some examples from the dataset.

While human filtering of such quantity is virtually impossible, automated attempts via CLIP models offer a low-cost filtering approach in exchange for a noisier dataset [3]. This inherent flaw in the dataset may lead to the trained model's dependency on prompts in the same format (internet prompt, unlike natural language) to provide enough textual information for accurate synthesis.

3. Technical Approach

3.1. Pre-trained Text-to-Image Model

We choose stable diffusion [1] because of its popularity, highly regarded performance, and open source availability.

(a) "animal_ears brown_hair cat_ears cat_tail japanese_clothes"

(b) "corporate stock certificate template 21 stock certificate templates free sle exle"

Figure 1. **LAION's Weakness**: examples of LAION prompts not written in natural human language

Stable diffusion is composed of 3 modules: the VAE, U-Net denoiser, and text encoder. The VAE is generally frozen at training time, while our experiments will alternate between training the text encoder and VAE. Stable diffusion version 2.1 is chosen to be the baseline for our experiments due to its recent release date and size, concerning our computational limitations.

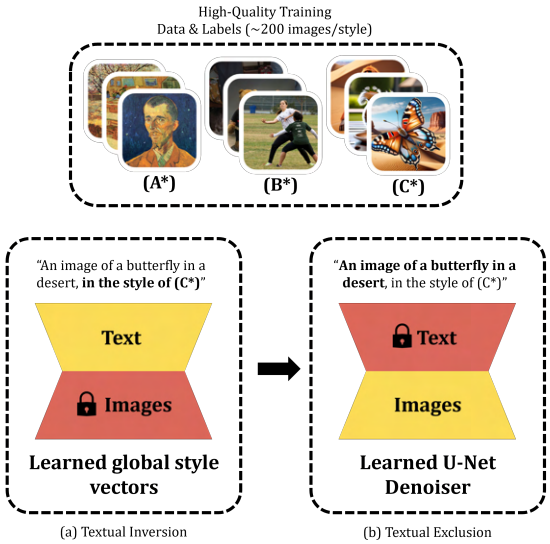
3.2. Textual Exclusion

Given a high-quality dataset $D = \mathcal{T} \times \mathcal{I}$ of text-image pairs with styles $I_1, I_2, \ldots, I_k \subseteq \mathcal{I}$, we aim to teach the text-to-image model $F : \mathcal{T} \mapsto \mathcal{I}$ to inherit the the text-image accuracy from the dataset. We follow the assumption that the dataset of all images $\mathcal{I}$ is a probabilistic mixture model of several components, including $I_1, I_2, \ldots, I_k$, each following a similar distribution (i.e. "A dog" remains a dog, despite being sampled from different styles). The text captions for each image are semantically accurate in natural human language, as this property is integral to ensuring our resulting model performs well on vague text prompts. This training process, described in Figure 2, can be separated into two integral steps, namely the textual inversion (3.2.1) and textual exclusion (3.2.2).

3.2.1 Learning Global Style Vectors

First, we employ textual inversion [5] to learn the individual styles from the training dataset. Textual inversion is a fine-tuning technique to learn specific objects or styles given a small amount of samples. Textual inversion randomly assigns a general prompt to the image, each equipped with an identifier token, which is initialized with a known word that most closely resembles the object/style in question (i.e. "A picture of a ⟨identifier⟩ dog", ⟨identifier⟩ is initialized with "corgi"). Every component of the diffusion model is frozen at training time, except for the newly added token.

Unlike the method presented in the original paper [5], however, we provide the text-to-image model with the accurate text-image pairs with style-specific tokens appended (i.e. "An image of a butterfly in a desert, in the style of ⟨C*⟩"). As we explicitly instruct the model with accurate textual descriptions for each image, the learned style vectors are less likely to be entangled with specific objects or entities commonly present in the dataset. These style tokens are initialized with zero vectors to eliminate inductive biases (further discussed in 4.1).



(a) Textual Inversion (b) Textual Exclusion

Figure 2. **Training Strategy**: Given a small and high-quality text-image dataset, we employ textual inversion to learn style-specific text embeddings. Subsequently, we train the U-Net denoiser conditioned on these embeddings to inherit the text-image accuracy from our dataset.

At training time, all model components are frozen except for the style vectors. This allows us to specifically learn the distributions of $I_1, I_2, \ldots, I_k$ while using said distributions to sample images from each style (i.e. $t \in \mathcal{T}$, $F(t|\text{from } I_1)$ generates an image from style #1).

3.2.2 U-Net Fine-tuning

Here, we aim to train the U-Net denoiser under our high-quality dataset. Each text-image pair is conditioned with styles learned from section 3.2.1, where we train F into F^* such that $F^*(t|\text{from } I_i)$ are accurate for $i = 1, 2, \ldots, k$. Assuming images are distributed similarly in each style, and that each style is independent of its textual and semantic contexts, learning $F^*(t|\text{from } I_i)$ can be generalized to learning $F^*(t)$ under any style. As real datasets don't generally follow this ideal assumption, using several styles should provide a remedy to its generalization capabilities.

This exclusion of style after fine-tuning allows the resulting model to retain its fidelity (similar to its pre-trained counterpart) while improving its generation quality and accuracy under text prompts, especially vague ones. Moreover, this improves the style-specific potency, allowing bet-

Figure 3. **Training Dataset**: each row corresponds to images sampled from a single dataset, namely MS-COCO, SDXL, and Van Gogh respectively.

ter representations of style-conditioned generations.

We employ Low-Rank Adaptation of Large Language Models (LoRA) [12] for its ability to perform on par with regular training with significantly lower computational cost. LoRA freezes all of the model's weights while applying low-rank decomposition differences to the weights, lowering the number of parameters.

3.3. Datasets

Prior research has shown that dataset quality plays a more important role than quantity in improving a model's performance [13]. As such, we collected three text-image datasets (Figure 3) with high text-image semantic correlation written in natural human language.

- **MS-COCO**: The dataset consisted of 328K images along with written captions in natural language. The selected images are photographed sceneries from the real world. This dataset reflects the style of realism.
- **SDXL**: We employ SDXL, the latest version of stable diffusion, to generate images conditioned on several object-scene prompts (i.e. "A butterfly in a desert") composed of 30 noun objects and 20 scenery. These prompts are appended (and omitted in textual exclusion) with 8-12 random adjectives related to photorealism, digital art, and aesthetics to ensure the generated images are of high quality. This dataset reflects the style of digital arts.
- **Van Gogh**: We also collected a relatively small dataset of Van Gogh art, where the pieces' names which reflect their semantic contents are used as text captions (i.e. "Portrait of a One-Eyed Man"). This dataset reflects the style of oil painting art.

Finally, 200 random samples from each dataset are gathered as training data.

4. Experiments

4.1. Vague Prompt Generation

To compare the accuracy of the generated images between our fine-tuned model and the original stable diffusion version 2.1, we use CLIP similarity scores [14]. Here, CLIP essentially measures how similar (or accurate in this case) the images are to their associated text captions. In particular, we compare the logits outputted from the CLIP [2] text and image encoders produced for a sample of 150 generated images per model using the same set of latent vectors. These images are generated with vague prompts from 30 different nouns (i.e. "A dog" or "A person"). Figure 4 illustrates the CLIP score logits between these models: Pre-trained stable diffusion version 2.1 and Textual exclusion (with and without initializer tokens for styles). These initial styles reflect each dataset's general feature: realistic, enhanced, and artistic for MS-COCO, SDXL, and Van Gogh respectively. Textual exclusion (with zero initializer tokens) outperforms the pre-trained stable diffusion version 2.1 significantly ($p = 0.0162$) in terms of generating images with vague prompts. Remarkably, several outliers are present in the similarity score of pre-trained stable diffusion version 2.1 – these correspond to failure cases we aim to overcome. Figure 5 depicts such images where pre-trained stable diffusion version 2.1 generations barely reflect their associated text prompts. Here, we observe the effects of textual exclusion enabling the model to generate more accurate results with vague prompts while retaining its synthetic diversity diversity.

We also demonstrate that style initialization with zero vectors achieves better performance, as hypothesized in section 3.2.1. This is because using known text tokens to initialize styles, while providing more visually appeasing results in multiple cases, introduces an inductive bias to how the styles are visually perceived by human standards. These biases prevent textual exclusion to separate style-specific features with text-image accuracy, hindering the resulting model's generalization capabilities.

4.2. Style-specific Generation

This experiment aims to test whether textual exclusion allows for generated images to resemble the original training data more than solely using textual inversion. Due to limited computational resources and training samples (only 200 per style), we were not able to provide enough images ($\geq 50,000$) to satisfy FID [15]'s requirements of a representative sample covariance. Hence, we chose KID, a similar measure proposed to facilitate the testing of the inception distance between datasets without restrictions on sample size [16]. Like the FID, the KID measures the distance between probability distributions. The KID utilizes maximum mean discrepancy (MMD) to measure the distance

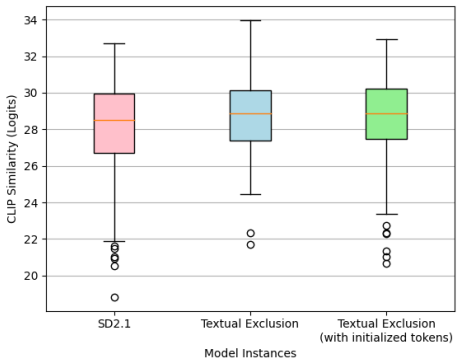


Figure 4. **Vague Prompt Similarity:** CLIP Score [14] logits between pre-trained stable diffusion [1] version 2.1 and textual exclusion (with and without style token initialization) of text-image pairs generated from vague prompts. Vanilla textual exclusion achieves the highest similarity score with a p-value of 0.0162 compared to the base model.

	Textual Inversion	Textual Exclusion
MS-COCO	0.0110	**0.0097**
SDXL	0.0216	**0.0199**
Van Gogh	0.0423	**0.0294**

Table 1. **Style-specific Generation:** KID [16] score computed on both models with each of the three datasets/styles.

between two probability distributions P and Q via the following equation [16]:

$$MMD(P, Q; \mathcal{H}) = \sup_{\mathcal{H}} \mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(Y)]$$

under an embedding space $\mathcal{H}$ and an embedding kernel f, essentially measuring the maximum distance of probability distributions after being embedded to known space. This eliminates the need to compute sample covariances, allowing for numerical stability while better reflecting the metric with a lower sample size.

In addition to enabling a higher text-image accuracy in a general case, textual exclusion is designed to enhance style-specific representation as both the U-Net and text encoder are fine-tuned towards the styles. Table 1 shows the measured result for each dataset. Textual exclusion's generations are more closely represented with each template style compared to textual inversion alone, proving textual inclusion (with style-specific instructions) a more potent version of textual inversion. Interestingly, textual inclusion outperforms by the largest margin on the Van Gogh dataset, which has a style that can be defined much clearer than the other two datasets.

4.3. Style Mixing

Textual exclusion learns a mixture model where, for every new dataset/style we performed textual exclusion upon, we have a new distribution learned specifically for that style. Extending from this idea, we explored the interpolation capabilities between each distribution. As we have explored the quality of our model in generating/modeling one of the style distributions in the previous section, we aim to use text-based conditioning (i.e. "in a style of ⟨A*⟩ and ⟨B*⟩ to interpolate between the styles we know.

We illustrate a comparison between individual styles and mixed styles in Figure 6. The first two rows depict textual exclusion's success in replicating (in a sense, overfitting to) the style of the fine-tuning datasets used (namely the SDXL and Van Gogh datasets). The third and fourth rows show style mixing attempts done by the model. Interestingly, the order of textual instruction governs the generated result to a certain extent. Particularly, the first style vector controls the global features of the images: the object's shape (i.e. the dog's breed), orientation (i.e. the car's direction), and accessories (i.e. hats). Subsequent style vectors allow for the images to be interpolated to other styles while retaining their general structure.

4.4. Subject-Driven Text-to-image Generation

Akin to textual inversion [5] and its successor DreamBooth [17], textual exclusion learns the distribution of an unseen feature via text guidance. As such, similar to textual exclusion's predecessors, we are interested in applying textual exclusion to the task described by both papers: subject-driven text-to-image generation. Given a small number of one template object, textual inversion and DreamBooth are designed to remember the appearance of said object and generate them in several contexts. As textual exclusion is shown to amplify the style representation from textual inversion (section 4.2), we hypothesize a similar result can be seen with subject-driven text-to-image generation.

Figure 7 depicts the three model's subject-driven generation given 5 images of the same dog under the prompt "A picture of a ⟨identifier⟩ dog on the moon". As described in DreamBooth's original paper [17], DreamBooth struggles with out-of-distribution generation. Though it excels at closely replicating the template dog, DreamBooth inadvertently fails at inserting the dog in an unknown environment (i.e. the moon). Textual inversion [5], on the other hand, fails to capture the dog's features while accurately generating the scenery of the moon, which we hypothesize is due to inheriting the pre-trained model's inability to generate out-of-distribution images. Textual exclusion, amplifying the model's knowledge of the dog by training the U-Net Model, performs outstandingly in the out-of-distribution generation with more accuracy and diversity than its predecessor techniques.

Figure 5. **Vague Prompt Image Generation:** Visual illustration of images generated under vague prompts using the same noise vectors. While pre-trained stable diffusion fails to generate the specified object in several cases, textual exclusion allows order to rise from the abstract chaos generated by the base model.

Figure 6. **Style Mixing**: Generated images conditioned on the style tokens “⟨SDXL⟩”, “⟨VANGOGH⟩”, “⟨SDXL⟩ and ⟨VANGOGH⟩”, and “⟨VANGOGH⟩ and ⟨SDXL⟩” (by row). From left to right, the prompts of each column are “a man”, “a dog”, “a car”, “a tower”, and “a room”, together with the specified style tokens.

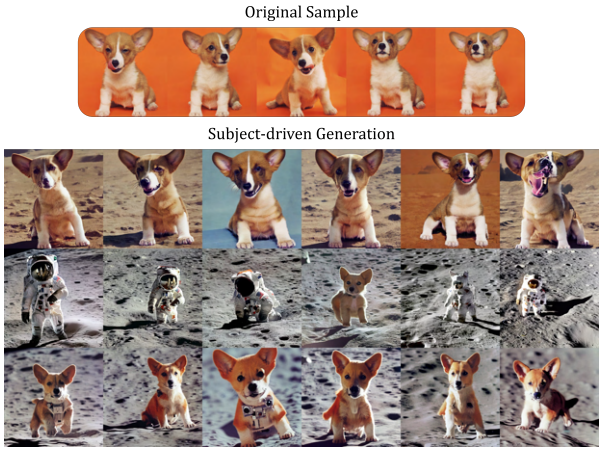


Figure 7. **Subject-Driven Text-to-Image Generation:** Generated images of original Dreambooth [17] (Row 1), textual inversion [5] (Row 2), and textual exclusion (Row 3) fine-tuned with 5 images of the same object. The generations are prompted with “A picture of a ⟨identifier⟩ dog on the moon”

5. Conclusion

5.1. Strengths of Textual Exclusion

We presented “textual exclusion”, a fine-tuning approach to text-to-image diffusion models that, given high-quality text-image labels of various styles, inherits the high text-image agreement while excluding style-specific feature, essentially generalizing the dataset quality given a small training dataset. Our idea is to train a special token to describe each style’s distribution, then train the denoising U-Net while exclusively conditioning the training captions on the style tokens. Remarkably, textual exclusion allows stable diffusion to generate more accurate images given vague prompts. Moreover, textual exclusion with specific style conditions is shown to be a more potent variant of textual inversion, generating a closer representative with the template styles while allowing styles to be mixed via textual instructions. We also applied textual exclusion to subject-driven generation and found that it performs better than DreamBooth and textual inversion at out-of-distribution generation.

5.2. Limitations

Despite requiring a small dataset, textual exclusion still requires an amount of high-quality image-text label pairs, meaning it is highly supervised. Our assumption that styles and text distributions are independent is highly ideal and unlikely to occur in real datasets.

5.3. Future Works

Textual exclusion is highly adaptable – only LoRA-trained weights and an extra text token are created for a textual exclusion instance. Textual exclusion could be used to improve the base performance of pre-trained text-to-image diffusion models for further research and recreational use. Moreover, its ability to be easily added to a pipeline allows for integration with other state-of-the-art diffusion techniques (i.e. our subject-driven generation combined with position-controllable techniques like GLIGEN [18] to generate any object in any position).

Some factors affecting textual exclusion's capacity, like the precise number of styles, have not been thoroughly studied yet.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 1, 2, 5

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *arXiv:2103.00020v1 [cs.CV]*, 2021. 1, 2, 4

[3] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25278–25294, 2022. 1, 2

[4] H. Zhuang, Y. Zhang, and S. Liu, "A pilot study of query-free adversarial attack against stable diffusion," *arXiv:2303.16378v2 [cs.CV]*, 2023. 1, 2

[5] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," 2022. 1, 3, 5, 6

[6] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, and M. Bennamoun, "Text to image synthesis for improved image captioning," *IEEE Access*, vol. 9, pp. 64918–64928, 2021. 2

[7] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," 2021. 2

[8] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022. 2

[9] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022. 2

[10] J. Wu, X. Gong, and Z. Zhang, "Self-supervised implicit attention: Guided attention by the model itself," 2022. 2

[11] S. Hong, G. Lee, W. Jang, and S. Kim, "Improving sample quality of diffusion models using self-attention guidance," 2023. 2

[12] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021. 4

[13] X. Dai, J. Hou, C.-Y. Ma, S. Tsai, J. Wang, R. Wang, P. Zhang, S. Vandenhende, X. Wang, A. Dubey, *et al.*, "Emu: Enhancing image generation models using photogenic needles in a haystack," *arXiv preprint arXiv:2309.15807*, 2023. 4

[14] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," *CoRR*, vol. abs/2104.08718, 2021. 4, 5

[15] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a nash equilibrium," *CoRR*, vol. abs/1706.08500, 2017. 4

[16] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018. 4, 5

[17] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," 2023. 5, 6

[18] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22511–22521, 2023. 7