Text-based Image Editing using Diffusion Model and CLIP

Apoorva Joshi
aj3215@columbia.edu

Rahul Aditya
ra3261@columbia.edu

Abstract

Image editing is a crucial aspect of various domains, yet traditional techniques often rely on manual intervention and struggle to incorporate textual descriptions seamlessly. In this study, we propose an automated text-guided image editing approach, which aims to replace specified objects while preserving the original context. Our method streamlines the editing process by enabling users to specify replacements through text prompts, eliminating the need for manual interventions. Our contributions encompass two key aspects. Firstly, we have developed an end-to-end pipeline that empowers users to specify objects or entities for editing and their replacements solely through textual descriptions. Secondly, we ensure the preservation of background context, a feature often absent in many text-guided image manipulation generative models. We achieve background preservation by incorporating L-2 and LPIPS [16] loss functions in the background, while employing a text-guided diffusion model using CLIP [7] for image editing – a capability lacking in many existing methods. To validate our approach, we compare it against DALLE2, a comparable model, demonstrating its effectiveness in seamlessly integrating textual concepts into image manipulation tasks while preserving visual coherence. Our work paves the way for new possibilities in creative expression and data augmentation for artists, designers, and researchers, offering a streamlined workflow.

[1]

1. Introduction

Image editing has emerged as an indispensable tool across diverse domains, encompassing creative arts and scientific research. However, conventional image editing techniques often necessitate manual intervention, such as manually indicating/scribbling specific regions for modification and identifying replacement objects or entities. Moreover, these methods frequently lack the seamless integration of textual descriptions into the editing process. While advancements like DALLE2 [10] have demonstrated the potential for incorporating textual prompts into image editing, they are not without their limitations. For instance, DALLE2 expects input image to be square-shaped of fixed dimensions (1024x1024), which is not necessary in our model. Additionally, many existing methods struggle to preserve the background context when replacing entities, often resulting in drastic alterations to the entire image. In light of these challenges, we propose an automated approach to address the task of text-guided image editing. By streamlining the editing process and leveraging advancements in machine learning and natural language processing, our method aims to seamlessly integrate textual descriptions into image manipulation tasks while preserving the contextual integrity of the original scene.

Our objective is to address the following challenge: when provided with an input image I, replacement prompt r, and edit prompt e, our goal is to generate an output image O. This output image should seamlessly replace the object or entity described in the replacement prompt r with the corresponding object or entity specified in the edit prompt e. Importantly, we aim to ensure that the remainder of the image I remains as unchanged as possible, preserving its original context and visual coherence. By enabling users to directly manipulate images based on textual descriptions, our proposed approach streamlines the image editing workflow and opens up new possibilities for creative expression. Artists, designers, and photographers can benefit from the ability to seamlessly integrate textual concepts into their visual compositions, while researchers can leverage the technique for data augmentation, image synthesis, and scene manipulation tasks.

We assess the effectiveness of our image editing approach by comparing it against text prompts using DALLE2 [10], given its similar capabilities and objectives to our method. Our contributions lie in two main aspects. Firstly, we have developed an end-to-end pipeline that empowers users to specify objects or entities for editing and their replacements solely through textual descriptions. Secondly, we ensure the preservation of background context, a feature often absent in many text-guided image manipulation generative models. We achieve background preservation

[1] https://drive.google.com/drive/folders/1o8PVp7Yhw3vuYHlM8S68kqCNi1HYQv6H?usp=drive_link

by incorporating L-2 and LPIPS [16] loss functions in the background, while employing a text-guided diffusion model using CLIP [7] for image editing – a capability lacking in many existing methods.

(a) r:Truck (b) mask (c) e:Camper RV

Figure 1. (a) depicts the input image I, (b) illustrates the mask generated by the segmentation model and CLIP with the edit prompt r = Truck, and (c) displays the manipulated image where the truck in (a) is replaced with an RV using the replacement prompt e = Camper RV.

2. Related Work

Numerous studies have explored text-based image manipulation using generative models such as GANs [2, 5]. Notably, DALL-E [11] introduced a multi-step approach involving a discrete VAE [12] to establish a compact context, subsequently fed into a transformer [13]. This transformer model is jointly trained on pairs of images and corresponding text prompts. Several variants of the DDPMs, such as GLIDE [9] and DALLE2 [10], have demonstrated state-of-the-art performance in text-to-image generation tasks. However, these models typically generate images based on textual inputs, operating on all pixels of the image. In contrast, our approach specifically targets the editing of entities matching text prompts, thereby offering a distinct focus on selective image manipulation.

Some models utilize a pretrained CLIP model [7] to guide generative models [3] in matching a text description. However, many of these approaches are primarily tailored for artistic image generation, often lacking natural realism characteristic of real-world images. In contrast, our method employs a similar technique to direct a diffusion model to edit images by replacing entities mentioned in the edit prompt with those described in the replacement prompt, while preserving the background as faithfully as possible. This approach ensures that the background of the resulting image maintains a natural appearance, with only the specified entities being replaced.

3. Methodology

Our aim is to generate an output image O from an input image I, a replacement prompt r, and an edit prompt e. In O, we aim to replace the object or entity described in r with the corresponding object or entity from e. It's crucial that the remainder of the image I remains unchanged to the greatest extent possible.

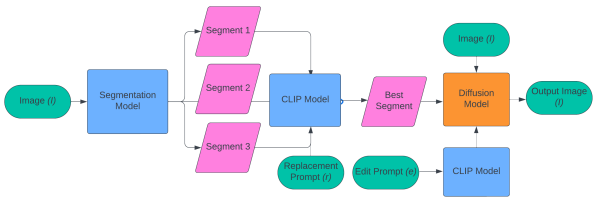
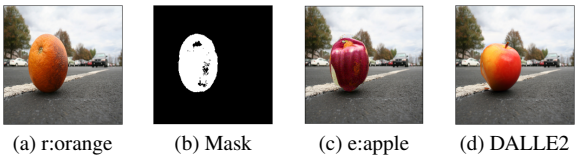


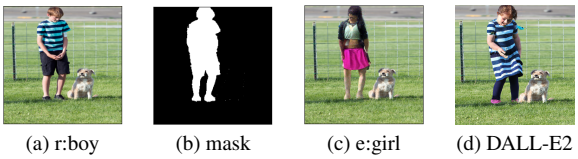
Figure 2. Flow Diagram of the Text-guided Diffusion Model for Image Editing



(a) r:orange (b) Mask (c) e:apple (d) DALLE2

(e) r:round (f) Mask (g) e:football (h) DALLE2

Figure 3. Each row presents an example with the input image shown in (a) alongside the replacement prompt r, the mask generated by the segmentation model depicted in (b), the edit prompt e shown in (c), and the resulting image produced by DALLE2.



(a) r:boy (b) mask (c) e:girl (d) DALL-E2

(e) r:dog (f) mask (g) e:cat (h) DALL-E2

Figure 4. Each row presents an example with the input image shown in (a) alongside the replacement prompt r, the mask generated by the segmentation model depicted in (b), the edit prompt e shown in (c), and the resulting image produced by DALLE2.

3.1. Semantic Image Segmentation

We employ the pretrained Semantic Image Segmentation model, specifically the variant nvidia/segformer-b5-finetuned-ade-640-640 [14], to generate masks corresponding to different segments within the input image I. These masks delineate distinct objects or entities present in the image.

We begin by preprocessing the input image I using the SegformerImageProcessor [14]. This preprocessing step in-

volves resizing the image and converting it into pixel values suitable for segmentation by the model. The Segformer model then generates segmentation logits, representing the likelihood of each pixel belonging to a particular object class. These logits are post-processed to obtain the final segmentation map S, which assigns a unique label to each pixel, indicating the object or entity it belongs to. The segmentation process results in the generation of a set of masks corresponding to different segments in the input image. These masks are denoted as $\{m_1, m_2, ..., m_k\}$, where each mask m_i represents a distinct object or entity within the image.

3.2. Using CLIP to pick the best mask

The replacement prompt r provided by the user is encoded using the CLIP model to obtain a textual feature vector CLIP(r) [7] that captures the semantic meaning of the prompt. Each segmented object can be represented as $s_i = m_i \odot I$ for all $i \in \{1, ..., k\}$ where $\odot$ represents the element-wise product. Each segmented object in the image is represented as an image embedding CLIP(s_i) by passing its corresponding mask through CLIP.

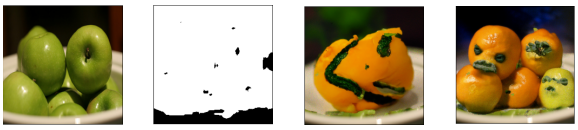
Following the encoding of the replacement prompt and the segmented objects, the next step involves computing probabilities to assess the relevance of each segment to the replacement task. This computation is based on the cosine similarity between the textual features extracted from the replacement prompt and the image embeddings obtained from the segmented objects.

By leveraging CLIP's ability to learn associations between images and corresponding textual descriptions, we compute the cosine similarity between each segment image embedding CLIP(s_i) and the text embedding CLIP(r). These cosine similarity scores are then passed through a softmax function to obtain probabilities. These probabilities reflect the likelihood of each segment aligning with the semantic context provided by the replacement prompt.

Once the probabilities are computed, the segment with the highest probability s^*, indicative of its close alignment with the replacement prompt, is selected as the best candidate for replacement. We find the optimal mask m^* corresponding to the segment s^* which we use for replacement using Diffusion model.

3.3. Editing using Diffusion model

Let the mask relevant to the replacement prompt r be m. In this phase, we take as input the original image I, the generated mask m, and the edit prompt e. The objective is to produce the final image O, which closely resembles the original image I outside the masked region m, while adhering closely to the description provided in the edit prompt e inside the masked region m. Given a data distribution $I_0 \sim q(I_0)$, a series of latent variables $I_1, I_2, ..., I_T$ are



(a) r:bunch of apples (b) mask generated (c) e:bunch of oranges (d) e:bunch of oranges

Figure 5. Figure presents an example with the input image shown in (a) alongside the replacement prompt r, the mask generated by the segmentation model depicted in (b), and result corresponding to two different edit prompts are shown in (c) and (d).

(a) r:bag (b) Mask (c) e:garbage

Figure 6. Figure presents an example with the input image shown in (a) alongside the replacement prompt r, the mask generated by the segmentation model depicted in (b), and the result is shown in (c).

generated by a forward Gaussian noise process with variance β_t at time t, where $\beta_t \in (0, 1)$. An important property demonstrated in the original DDPM paper [4] is the ability to directly sample I_t given I_0 without the need to generate intermediate latents. This is achieved by the equation:

$$I_t = \sqrt{\bar{\alpha}_t} I_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$, $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$.

We utilize CLIP once again to guide the diffusion model in generating latents that closely adhere to the edit prompt e. Since CLIP has been trained on images without Gaussian noise, it is reasonable to estimate the clean image from which we initially started. This estimation aids in computing the cosine distance, which serves as a term in the loss function, between the text embedding obtained by passing the edit prompt e and the image embedding of the estimated clean image $\hat{I}_0$ given I_t. We can utilize the previously derived equation to calculate $\hat{I}_0$ as follows:

$$\hat{I}_0 = \frac{I_t - \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon}{\sqrt{\bar{\alpha}_t}}$$

In order to confine the modifications made by the Diffusion model exclusively to the pixels encompassed by the input mask m, we employ cosine distance computation via CLIP on the embeddings of the estimated initial image within the mask, designated as $\hat{I}_0 \odot m$, and the text prompt e. Consequently, the CLIP loss function is formulated as the cosine distance between $CLIP(e)$ and $CLIP(\hat{I}_0 \odot m)$. This loss is denoted as L_{CLIP}.

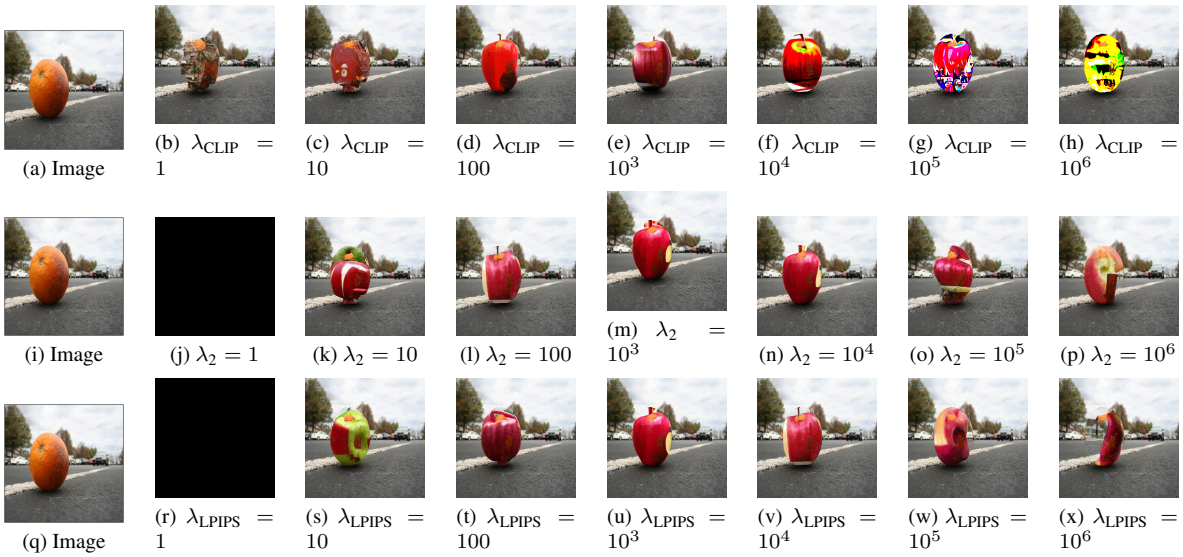


Figure 7. Ablation Study Results on hyperparameters λ_{CLIP}, λ_2 and λ_{LPIPS}. The first row shows the sweep of hyperparameter λ_{CLIP}, second row for λ_2 and the third one for λ_{LPIPS}

.

Observations reveal that while L_{CLIP} is computed exclusively for pixels within the mask, the influence extends to pixels outside the mask. This occurrence stems from the initialization using an isotropic Gaussian noise and the subsequent denoising process lacking background constraints. To address this issue and preserve the background, we introduce two additional loss functions.

Firstly, we define the background image as $I_b = I \odot (1 - m)$ and the foreground image to be edited as $I_f = I \odot m$. To maintain the integrity of the background I_b, we employ the L_2 norm distance between the images $\hat{I}_0 \odot (1 - m)$ and I_b. This entails computing the mean squared error over the pixel values of the two images. Additionally, we incorporate the Learned Perceptual Image Patch Similarity metric (LPIPS) [16] between $\hat{I}_0 \odot (1 - m)$ and I_b in the loss function to further ensure the preservation of the background.

Let the defined loss functions corresponding to the background preservation measures be denoted as L_2 and L_{LPIPS}. Consequently, the composite loss function is formulated as $\lambda_{\text{CLIP}} L_{\text{CLIP}} + \lambda_{\text{LPIPS}} L_{\text{LPIPS}} + \lambda_2 L_2$, where hyperparameters $\lambda_{\text{CLIP}}, \lambda_{\text{LPIPS}}$ and λ_2 are introduced. The results section includes an ablation study of these hyperparameters. All experiments were conducted using images from the COCO dataset [8] and the Open Images Dataset v4 [6], with replacement prompt r and edit prompt e generated by us. These prompts were designed to pose challenges or provide crucial insights for the ablation study.

4. Results

In our diffusion model, we incorporated three hyperparameters: λ_{CLIP}, λ_{LPIPS}, and λ_2. To optimize the performance of our model, we conducted a series of hyperparameter tuning experiments, varying each hyperparameter across orders of magnitude from 1 to 10^6. Our qualitative assessments indicate that the generated images align well with our image editing task when setting each hyperparameter to 10^3. Detailed ablation studies for each hyperparameter are presented in Sections 4.2.2, 4.2.3, and 4.2.4.

4.1. Image editing and Comparison with DALLE2

In this section, we conduct a comparative analysis between the outcomes of our model and those produced by DALLE2 [10], serving as a benchmark for evaluating performance. Figure 3 showcases the comparison of replacement prompts r for the same segment, juxtaposed against DALLE2's results. The edit prompt e is fixed as $orange$, with r set as $apple$ for the first experiment and $football$ for the second. In Figure 4, we explore different segments of the identical input image, varying the edit prompt to boy in one experiment and dog in the other. Consequently, distinct masks are generated, and the outcomes are benchmarked against DALLE2, with r set as $girl$ in the first case and cat in the second. Across all scenarios, our results exhibit comparable performance to DALLE2, with a notable emphasis on minimal changes to the background and no prerequisite image preprocessing, as required by DALLE2 (1024x1024). Moreover, we observe discrepancies in the utilization of color as a guiding parameter be-

tween DALLE2 and our model, with DALLE2 focusing on color preservation.

4.2. Ablation Study

In all the ablation study experiments presented in the next sections, we fix the input image as shown in the Figure 7 with the edit prompt $e = orange$ and the replacement prompt $r = apple$. In the following experiments, we demonstrate how the introduced loss functions contribute to solving the problem at hand. Notably, L_{CLIP} ensures that the edited image aligns with the edit prompt e, while L_{LPIPS} and L_2 losses are instrumental in preserving the background. This work closely follows the work done by Bau et. al [1].

4.2.1 Ablation Study on λ_{CLIP} to match edit prompt

In this ablation study, we fix the values of $\lambda_{\text{LPIPS}} = 1000$ and $\lambda_2 = 1000$, while varying the value of λ_{CLIP} over the range $1, 100, 1000, 10^4, 10^5$, and 10^6. This hyperparameter controls the degree to which the resulting image resembles the replacement prompt r. We observe that for lower values of λ_{LPIPS} (e.g., 1, 10), the edited image resembles an orange with a mixed background, rather than an apple. Conversely, for higher values of λ_{LPIPS} (such as $10^4, 10^5$, and 10^6), the coloration of the images becomes excessively vivid, deviating from the typical appearance of an apple.

The ablation experiment highlights the significance of λ_{CLIP} in aligning the image edits with the edit prompt e. Lower values often result in a failure to produce the desired output, while larger values tend to produce exaggerated edits. The optimal balance is achieved at $\lambda_{\text{CLIP}} = 1000$.

4.2.2 Ablation Study on λ_2 and λ_{LPIPS} to preserve background

For ablation study on λ_2, we fix the values of $\lambda_{\text{CLIP}} = 1000$ and $\lambda_{\text{LPIPS}} = 1000$, while varying the value of λ_2 over the range $1, 100, 1000, 10^4, 10^5$, and 10^6. Similarly for ablation study on λ_{LPIPS}, we fix the values of $\lambda_{\text{CLIP}} = 1000$ and $\lambda_2 = 1000$, while varying the value of λ_{LPIPS} over the range $1, 100, 1000, 10^4, 10^5$, and 10^6.

In both cases, when the values of λ_2 and λ_{LPIPS} are set to 0, the resulting image is entirely black. Conversely, for higher values of λ_2 and λ_{LPIPS} (such as 10^5 and 10^6), the output image extends beyond the masked region, aiming to preserve the background as faithfully as possible. These ablation studies underscore the significance of λ_2 and λ_{LPIPS} in maintaining the integrity of the background in the input image outside the mask.

4.3. Shortcomings

A notable limitation arises in obtaining suitable mask images from the segmentation model. Since we utilize a pretrained segmentation model for inference, the generated segments may not be fully optimized for CLIP to select the most appropriate one. This issue becomes particularly evident when the replacement prompt spans multiple objects, as illustrated in Figure 5, or when dealing with grayscale images, as depicted in Figure 6. Furthermore, when the edit prompt involves multiple objects, the resulting image may not contain all the specified entities, as seen in Figure 5(c). Even when multiple objects are generated, they may not align precisely with the desired entities, as demonstrated in Figure 5(d). Additionally, in the grayscale image shown in Figure 6, the diffusion model introduces color into the background, compromising the overall image's fidelity.

5. Conclusion

In conclusion, our work introduces an automated text-guided image editing approach that efficiently replaces specified objects while preserving the original context, a departure from conventional text-guided generative models that often fail to preserve the background. Through ablation studies, we demonstrated the necessity of the introduced loss functions in achieving this task. However, there remain open challenges and opportunities for future research. One limitation of our approach is its reliance on pre-trained segmentation models, which may produce inaccurate masks. Enhancing the accuracy of segmentation models or exploring alternative methods for mask generation could bolster the robustness of our approach. Additionally, while our method currently focuses on replacing single objects or entities specified by text prompts, extending it to handle more complex editing tasks, such as multiple object replacements or semantic edits, could enhance its versatility. We can also introduce image augmentations to make it robust against adversarial attacks. Future work could also involve experimenting with CLIP for textual-guided segmentation, leveraging color as a guiding parameter in textual prompts, and exploring alternative generative models beyond DDPM [4] like DDIM [15]. These avenues hold the potential to further advance the capabilities of text-guided image editing systems and empower users with more flexible and intuitive editing tools.

6. Contribution

The work was equally divided in terms of literature review, drawing key-insights, implementation, and report write-up. Image Editing, Benchmarking with DALLE2 [10], and ablation study on λ_{CLIP} was performed by Apoorva and the ablation study on λ_2 and λ_{LPIPS} was performed by Rahul.

References

[1] Alex Andonian, Sabrina Osmany, Audrey Cui, YeonHwan Park, Ali Jahanian, Antonio Torralba, and David Bau. Paint

by word. *arXiv preprint arXiv:2103.10951*, 2021. 5

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 5

[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[6] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020. 4

[7] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 1, 2, 3

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 4

[9] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021. 2

[10] 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. 1, 2, 4, 5

[11] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2

[12] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[13] A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. Attention is all you need. In *NIPS*, 2017. 2

[14] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021. 2

[15] Qinsheng Zhang, Molei Tao, and Yongxin Chen. gddim: Generalized denoising diffusion implicit models. *arXiv preprint arXiv:2206.05564*, 2022. 5

[16] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1, 2, 4