ReEdit: Multimodal Exemplar-Based Image Editing

Ashutosh Srivastava[1*†] Tarun Ram Menta[2*] Abhinav Java[3*‡] Avadhoot Jadhav[4†]
Silky Singh[5‡] Surgan Jandial[6‡] Balaji Krishnamurthy[2]
[1] Indian Institute of Technology, Roorkee [2] Adobe MDSR [3] Microsoft Research
[4] Indian Institute of Technology, Bombay [5] Stanford University [6] Carnegie Mellon University

Abstract

Modern Text-to-Image (T2I) Diffusion models have revolutionized image editing by enabling the generation of high-quality photorealistic images. While the de-facto method for performing edits with T2I models is through text instructions, this approach is non-trivial due to the complex many-to-many mapping between natural language and images. In this work, we address exemplar-based image editing – the task of transferring an edit from an exemplar pair to a content image(s). We propose ReEdit*, a modular and efficient end-to-end framework that captures edits in both text and image modalities while ensuring the fidelity of the edited image. We validate the effectiveness of* ReEdit *through extensive comparisons with state-of-the-art baselines and sensitivity analyses of key design choices. Our results demonstrate that* ReEdit *consistently outperforms contemporary approaches both qualitatively and quantitatively. Additionally,* ReEdit *boasts high practical applicability, as it does not require any task-specific optimization and is* 4× *faster than the existing state-of-the-art. The code and data for our work is available at https://reedit-diffusion.github.io/.*

1. Introduction

Image editing [1, 11, 21, 30, 35] is a rapidly growing research area, with a wide range of practical applicability in domains like multimedia, cinema, advertising, etc. Recent advancements in text-based diffusion models [17,35,37,43] have accelerated the progress in the field of image editing, yet diffusion models remain limited in their practical viability to real world applications. For example, if a practitioner is making detailed edits—such as transforming a scene from daytime to nighttime—and wants to apply the same adjustments to multiple images, they would face a considerable challenge, since crafting each image individually can be time consuming. In such cases, simple textual prompts might not be sufficient to achieve the desired consistency and efficiency.

Notably, an ideal editing application should be **fast**, have the ability to understand the exact **user intent** and produce **high fidelity** outputs. Most existing work in this domain leverages textual descriptions to perform image editing [4, 14, 19, 22, 23, 33, 36, 49], however, text is inherently limited in its ability to adequately describe edits. These challenges motivate us to focus on a relatively unexplored field of *exemplar based image editing*. This formulation is motivated by 'visual prompting' proposed by Bar et al. [3].

Existing works in this area typically optimize a text embedding during inference to capture each edit [20,34] which is time taking. Other methods like [15,48] utilize sophisticated models trained specifically for the task of editing like InstructPix2Pix [4] (IP2P), which requires a large labelled training dataset. These datasets can be extremely difficult to obtain due to the nature of the problem. Further, recent approaches like VISII [34] can only capture a limited type of edits (performs well only for *global style transfer* type edits) as a result of the way its text embedding is optimized.

Unlike existing approaches, we propose an **efficient** end-to-end **optimization-free** framework for exemplar based image editing - **ReEdit**. The proposed framework consists of three primarily components - *first* we capture the edit from the exemplar in the image embedding space using pretrained adapter modules [41], *second*, we capture the edit in natural language by incorporating multimodal VLMs like [28] capable of detailed reasoning, and *last* we ensure that the content and structure of the test image is maintained and only the relevant parts are edited by conditioning the image generator on the features and self attention maps [49] of the test image. To summarize, the contributions of our work are listed below:

1. We propose ReEdit, an inference-time approach for *exemplar-based image editing* that does not require any model fine-tuning or inference time optimization. Compared to the existing state-of-the-art, the runtime of method is ∼4x faster, and is independent of the base diffusion model.

[*] Equal Contribution
[†] Work done during internship at Adobe MDSR
[‡] Work done while at Adobe

2. We collate a dataset of 1500 exemplar pairs (x, x_{edit}), and corresponding test images with ground truth (y, y_{edit}), covering a wide range of edits. Due to a lack of standardized datasets, our dataset paves towards a standardized evaluation of *exemplar-based image editing* approaches.
3. Our rigorous qualitative and quantitative analysis shows that our method performs well on a variety of edits while preserving the structure of the original image. These observations are corroborated by significant improvements in quantitative scores over baselines.

2. Related Work

Diffusion Models. Prior to diffusion models, GANs [13, 59, 63] were the de-facto generative models used for (conditional) image synthesis and editing. However, training GAN networks is prone to instability and mode collapse, among many issues. Recently, large-scale text-to-image generative models [8, 9, 39–41, 43, 58] have benefitted from superior model architectures [50] and large-scale training data available on the internet. Of particular interest is diffusion models [17, 35, 37, 40, 41, 43, 46], that are trained to denoise random gaussian noise resulting in high-fidelity and highly diverse images. These models are typically trained on millions of text-image pairs. In this work, we use a pretrained Stable Diffusion [41] model which operates in the latent space instead of the image pixel space.

Multimodal Vision-Language Models (VLMs). Multimodal VLMs [25, 27–29, 38, 45] have the remarkable capability to understand and process both texts and images. Two particularly useful works fall in the scope of this paper: CLIP [38] and LLaVA [27–29]. CLIP represents both images and texts in a shared embedding space. It was trained on 400M image-text pairs in a contrastive manner – maximizing the similarity between related image-text embeddings, while minimizing the similarity between unrelated image-text embeddings. LLaVA combines a visual encoder with Vicuna [5] to provide powerful language and visual comprehension capabilities. It has impressive capacity to follow user instructions based on visual cues.

Text-based Image Editing. Diffusion models, with their impressive generative capabilites, have also been adapted for image editing [7, 10, 19, 23, 24, 33, 36, 42, 49, 56, 60, 62]. Multimodal models like CLIP [38], and cross-attention mechanisms [50] have enabled conditioning a diffusion model to directly edit an image with a text input [2, 35]. SDEdit [32] takes an image as input along with a user guide, and subsequently denoises it using SDE prior to increase its realism. Other related works [6, 47] guide the generative process conditioned on some user input, for e.g., a reference image. Imagic [22] finetunes a diffusion model on a single image to perform image editing. Prompt-to-prompt [14] attempts to edit an image while preserving its structure by modifying the attention maps in a pretrained diffusion model. Similarly, pix2pix-zero [36] preserves the content and structure of the original image while editing via cross-attention guidance. Instruct-pix2pix [4] first collected a huge dataset of (image, edit text, edited image) triplets, and trained a diffusion model to follow edit instructions provided by a user. Plug-and-Play [49] aims to preserve the semantic layout of an image during an edit by manipulating spatial features and self-attention in a pretrained text-to-image diffusion model. Although these approaches produce plausible edits to an image, there still exist limitations where either the edit instruction/text is completely ignored, or the structure of the original image is drastically modified. Additionally, our work differs from this line of work since we get rid of text-based instructions altogether.

Exemplar-based Image Editing. In the field of Computer Vision, 'visual prompting' was first proposed by [3]. Later works [51, 52] built a generalist model based on visual in-context learning to solve multiple vision tasks, including segmentation. Exemplar-based editing methods [15, 20, 34, 48, 54] are an extension of ”visual prompting”, where the focus is to edit an image conditioned on a visual input, called *exemplar*. This can include insertion of the exemplar object in a given image to produce a photo-realistic output as in Paint-by-Example [54], or transfer of overall style from an exemplar image to a given image [20]. The concept of *image analogies* was proposed in [15] and later used in [26] for visual attribute transfer from one image to another, for e.g., color, tone, texture, style. It has also been extended for example-based editing using diffusion models [48]. The present work is closest to VISII [34] and ImageBrush [55] – both explore the idea of using an exemplar pair as visual instruction for image editing. Unlike our work, VISII [34] relies on optimization-based inversion to capture the edit in CLIP [38] text space, while ImageBrush [55] benefits from training a diffusion model on the revised conditional inpainting task. At the same time, our proposed approach achieves superior edit quality without the need for any optimization or additional training.

3. Methodology

In this section, we first introduce some preliminaries and describe the notation. We then introduce our proposed framework, ReEdit that comprises two key steps: **(a)** capturing the edit (g) from the given pair of exemplars in both text and image space, followed by **(b)** conditioning the diffusion model (M) to apply this edit on a test image (y) without any optimization. The overview of our framework is illustrated in Fig. 1.

Problem Setting and Notation. Given a pair of exemplar

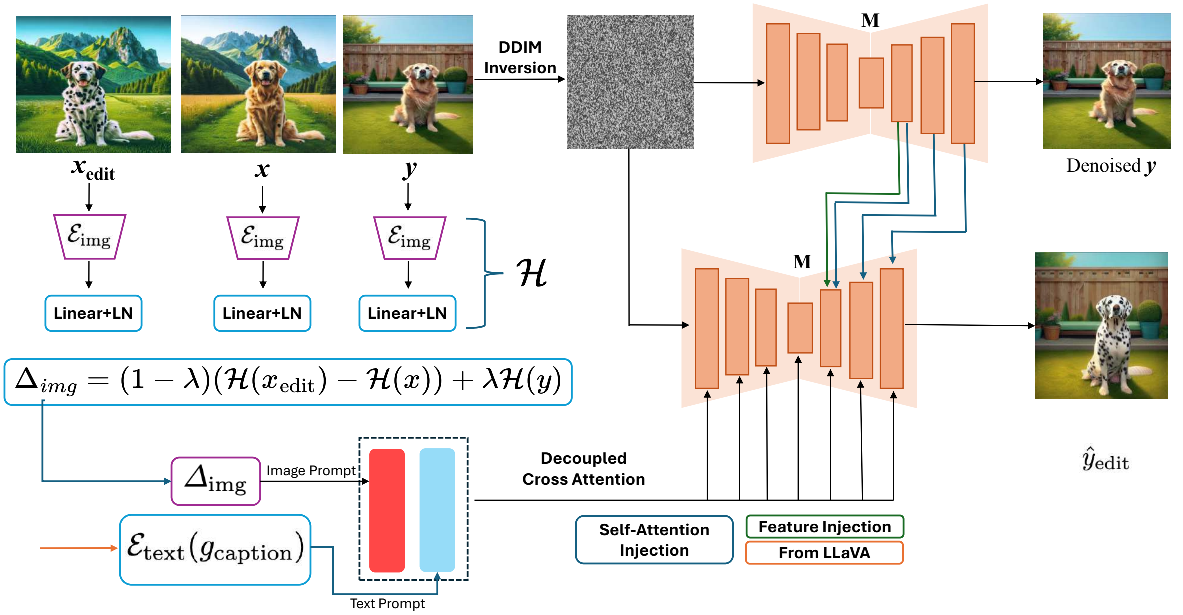
DDIM Inversion
M
Denoised y
x_{edit}
x
y
$\mathcal{E}_{\text{img}}$
$\mathcal{E}_{\text{img}}$
$\mathcal{E}_{\text{img}}$
$\mathcal{H}$
Linear+LN
Linear+LN
Linear+LN
M
$\Delta_{img} = (1-\lambda)(\mathcal{H}(x_{\text{edit}}) - \mathcal{H}(x)) + \lambda\mathcal{H}(y)$
Δ_{img}
Image Prompt
Decoupled Cross Attention
$\hat{y}_{\text{edit}}$
$\mathcal{E}_{\text{text}}(g_{\text{caption}})$
Self-Attention Injection
Feature Injection
From LLaVA
Text Prompt

Figure 1. Overview of our framework ReEdit. ReEdit identifies the required edit information from a pair of exemplar images (x, x_{edit}) in both text and image spaces, and uses a diffusion model to apply this edit on a new target image y. For additional details are provided in Section 3 and Appendix A.1. Best viewed in color.

images (x, x_{edit}), where x denotes the original image, and x_{edit} denotes the edited image respectively. Our objective is to capture the edit (say g, such that $x_{\text{edit}} = g(x)$), and apply the *same edit* (g) on a test image y to obtain the corresponding edited image $\hat{y}_{\text{edit}}$. Let $M(\theta)$ denote a pre-trained diffusion model (here, SD1.5 [41]) parameterized by θ, where θ remains frozen. And let $\mathcal{E}_{\text{img}}$ and $\mathcal{E}_{\text{text}}$ denote pre-trained CLIP image and text encoders respectively.

Background. Recent work [57] proposes to utilize simple adapter modules to generate high quality images with images as prompts. Unlike typical T2I models whose cross attention parameters are only conditioned on text-embeddings, IP-Adapter [57] adds newly initialized linear and cross attention layers and finetunes these additional parameters (∼22M), which directly allow the introduction of image embeddings to pretrained T2I models. As motivated in Sec 1, text alone often falls short in capturing the edit from exemplar pairs, so we propose a strategy that enables us to capture the edits from the exemplar pairs both in the image space (using simple adapters) and in the text space.

3.1. Capturing Edits from exemplars

We posit that *textual descriptions are necessary but not sufficient* to generate $\hat{y}_{\text{edit}}$ from (x, x_{edit}, y). Consequently, we capture edits in both *text* and *image* space.

Edits in natural language. Firstly, we leverage a multimodal VLM (LLaVA [27–29]) to verbalize the edits in the exemplar pair $(x,\ x_{\text{edit}})$. We pass these images as a grid, along with a detailed prompt p_1 that instructs LLaVA to generate a complete description of the edits, denoted by g_{text}. Additionally, to provide the context of the test image y, we curate another prompt p_2 instructing LLaVA to describe $\hat{y}_{\text{edit}}$ in text after applying the edit g_{text} on y. As a result, we obtain a final text description of $\hat{y}_{\text{edit}}$, denoted by g_{caption}. To reduce verbosity and token length, we limit g_{caption} to 40 words. Refer to Appendix A.1 for the exact prompts p_1 and p_2, and an overview of the caption generation process.

Edits in image space. Natural language cannot capture the specific style, intensity, hue, saturation, exact shapes, or other nuances in the image. Therefore, we also capture the edits from (x, x_{edit}) and the original image (y) directly in the CLIP embedding space. Specifically, we apply a pre-trained linear layer and layer norm [57] on the embeddings of x, x_{edit} to make the embeddings compatible with M. The edit is captured as follows - $\Delta_{img} = \lambda(\mathcal{H}(x_{\text{edit}}) - \mathcal{H}(x)) + (1-\lambda)\mathcal{H}(y)$; where $\mathcal{H}(x) = \text{LN}(\text{Lin}(\mathcal{E}_{\text{img}}(x)))$ and LN, Lin are the layer norm and linear projection operators respectively. The **edit weight** slider is denoted by λ, that weighs the contributions of the edit and the target image while generating the final result y_{edit}. Our final edit embedding is hence given by the pair $g := (\Delta_{\text{img}},\ \mathcal{E}_{\text{text}}(g_{\text{caption}}))$. Both the image and text conditioning in g work in tandem to provide nuanced guidance for precise edits. As shown in Fig 1, the edit embeddings in g are processed by their respective decoupled cross attention parameters and propagated through M to generate the final image.

3.2. Conditioning Stable Diffusion on (g, y)

A crucial requirement of image editing approaches is that they preserve the content and structure of the original image in the edited output. Thus, we aim to condition M on g such that only the relevant parts of y are edited, while the rest of the image remains intact. To achieve this, we introduce the approach of attention and feature injection motivated by [49]. Specifically, we invert y using DDIM inversion [46], and run vanilla, unconditioned denoising on the inverted noise (y_{noise}). During this process, the intermediate features (f) and attention matrices (Q, K) are extracted from the up-sampling blocks, and these features contain the overall structure information for y [49]. Finally, to generate the final edited image, we start with y_{noise}, and condition the denoising process on the edit g (through cross-attention), inject the features (f) at the fourth layer, and modify the keys and queries (Q, K) in the self-attention layers from layers 4 to 11 of M to obtain $\hat{y}_{\text{edit}}$.

4. Dataset Creation

Our method is an inference-time approach, and is directly applicable to an arbitrary set of (x, x_{edit}, y) images. However, there are no existing evaluation datasets for exemplar-based image editing in the current literature. Hence, we curate a dataset from the existing image editing dataset. Specifically, the exemplar pairs are taken from the InstructPix2Pix dataset. This is a dataset for text-based image editing, containing $450,000$ $(x, x_{\text{edit}}, g_{\text{edit}})$, where x_{edit} is the image obtained after applying the edit instruction g_{edit} on input image x. This dataset was generated by applying Prompt-to-Prompt [14] on a Stable Diffusion model. We found two common issues with this dataset - i) the edit pair $(x, x_{\text{edit}}$ did not adhere to the edit instruction g_{edit}, and ii) The edit instruction g_{edit} did not apply to the input image x. Refer to Fig. 2 for examples of these failure cases. As a result, we carefully curate a dataset of $(x, x_{\text{edit}}, y, y_{\text{edit}})$ where x, y, and the corresponding edited images are taken from IP2P samples with the same edit instruction g_{edit}. Through visual inspection, we manually ensure that the two aforementioned issues do not creep into our dataset, resulting in a high-quality dataset of $\sim$ 1500 samples, across a diverse set of edit types. We provide the exact statistics of our dataset, including the different types of edits in Table 1

Type of Edit	Number of Examples
Global Style Transfer	428
Background Change	212
Localized Style Transfer	290
Object Replacement	366
Motion Edit	14
Object Insertion	164
Total	**1474**

Table 1. Summary and statistics of the types of edits in the evaluation dataset. Special care was taken to ensure diversity of edit categories.

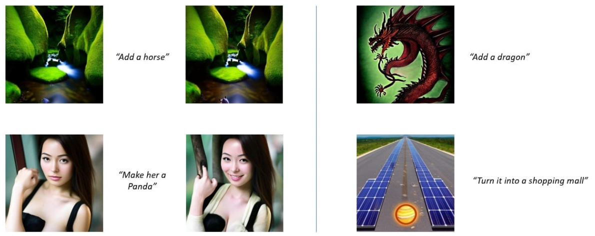

Figure 2. Examples of ambiguous samples in the InstructPix2Pix dataset, motivating the need for manual curation. Additional examples can be found in Appendix A.4

5. Experiments and Results

In this section, we first provide a detailed description of the implementation which includes the hyperparameter choices of both ReEdit and baselines. Next, equipped with our curated dataset, we evaluate the performance of ReEdit. The key feature of our dataset is the presence of a ground truth edited image, denoted by y_{edit} which enables us to use several standard image quality evaluation metrics. As a result, we compute five quantitative metrics across the full dataset. measuring structural and perceptual similarity between $(\hat{y}_{\text{edit}}, y_{\text{edit}})$ (LPIPS [61], SSIM [53]), faithfulness of the edit with the exemplar pair (CLIP score [16], Dir. Similarity [12], and S-Visual [34]). For each method in the comparison, the quantitative scores are reported by selecting the hyperparameter which yields the best average performance across all metrics. However, different edit types may require different hyperparamters to yield the best quality results, so in our qualitative analysis we choose the best per-sample hyperparameter for each method.

We present our quantitative results in Table 2, show several qualitative examples in Fig 3, and show the running times of all the methods in Table 3. Further, we also illustrate additional qualitative examples in Appendix A.3. A detailed discussion of our results can be found in Sec. 5.2. Appendix A.2 includes further details on the usage and implementation of the various metrics.

5.1. Implementation details

We carry out an extensive comparison of our framework ReEdit with existing approaches. The main work we compare against is VISII [34], another inference-time exemplar

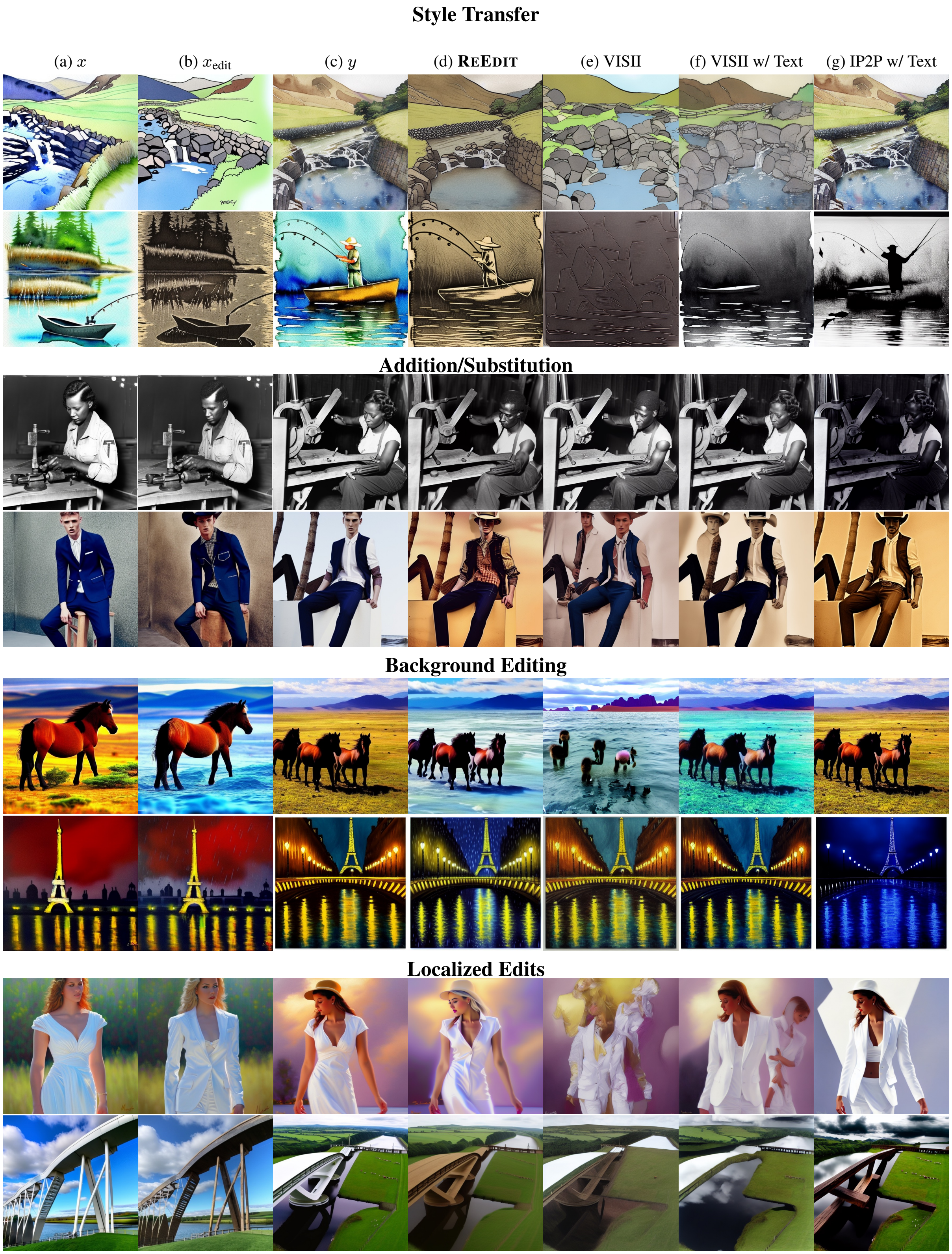

Figure 3. Qualitative comparison of our framework ReEdit with strong baselines (VISII, InstructPix2Pix) for exemplar-based image editing. ReEdit consistently produces images with higher edit accuracy and better consistency in non-edited regions compared to the baselines. Zoom in for better view. Additional results presented in Appendix A.3

based editing method. To ensure a fair comparison, we further augment VISII with LLaVA generated instructions as an additional baseline. Finally, we also compare against a text-based editing approach - InstructPix2Pix, once again using the LLaVA generated edit instruction as input. All baselines are evaluated on a single A100 GPU with 80GB memory, and all images are resized to 512×512 pixels. We now further describe the exact setup and hyperparameters for REEDIT and the various baselines when generating results for qualitative and quantitative analysis. We employ LLaVA-1.6 [28] for obtaining automated captions and edit instructions in all methods.

REEDIT. We use SD1.5 with IP-Adapter [57] as the base model. To compute the image prompt embedding CLIP ViT-L/14 is used, which is then pooled to 4 tokens. 3 images x, x_{edit} and y are used to generate the image prompt embedding Δ_{img}, and the previously describe pipeline to generate the text caption using LLaVA. The input to a standard text-to-image diffusion model (here, Stable Diffusion v1.5) is typically a sequence of 77 tokens (sequence length of CLIP text encoder), which after adding the image prompt becomes 81 tokens. The last 4 tokens are are separately processed in IP-Adapter's cross attention modules.

We perform DDIM inversion of the test image y for 1000 steps to generate y_{noise} for feature and self attention guidance. The features and self-attention (f, Q, K) from the vanilla denoising of y_{noise} are injected at each of the $4-11^{th}$ layers of the upsampling blocks respectively. An important parameter is the classifier free guidance (CFG [18]) weight. We fix the CFG to 10 across all experiemtns, and only vary the edit weight (λ) in our experiments as the sole hyperparameter in REEDIT.

a. VISII. [34] optimizes an edit instruction c_T in the latent space of the CLIP text encoder of InstructPix2Pix to learn the edit from the exemplar pair $(x, x_{\text{edit}}$. This learnt instruction c_T is used as input along with the test image y to obtain the desired edit. We adopt their original setup, and optimize for $T = 1000$ steps using AdamW [31] with learning rate 1e-4, and the respective weights for the loss term in VISII as $\lambda_{\text{mse}} = 4$ and $\lambda_{\text{clip}} = 0.1$. Following the original setup, for each inference, we perform 8 independent optimizations with different random seeds, and choose the c_T that minimizes the overall loss. We experiment with multiple values for text guidance from $8, 10, 12$ and use the default image guidance of 1.5 as the hyperparameters.

b. VISII with text. We introduce an important augmentation to VISII, concatenating an additional textual instruction as suggested in the original work. This helps guide the model using both natural language and image differences. We generate the edit text similar to the approach in Sec. 3. First, we pass a grid of exemplar pairs (x, x_{edit}) and a detailed prompt $p1$ to LLaVA to generate a detailed edit text g_{edit}. Next, instead of generating g_{caption} from $(g_{\text{edit}}, y, p2)$ we instead instruct LLaVA to generate a short summary of the *edit instruction* (say $g_{\text{edit-inst}}$) using $g_{\text{edit}}, y, p3$ where $p3$ is a simple modification of $p2$ instructing LLaVA to generate an edit instruction for InstructPix2Pix. Refer to Appendix A.1 for all prompts. The hyperparameters and optimization are the exact same as used for VISII.

c. InstructPix2Pix. We directly use a supervised pre-trained text-based image editing model, InstructPix2Pix [4]. Thought InstructPix2Pix was intended to be used with custom text instructions, the goal in this setting is to transfer edits without explicit supervision. Hence, we utilize LLaVA edit instructions $g_{\text{edit-inst}}$ as described in *b. VISII with text*. We experiment over the same set of hyperparameter values as the previous two baselines.

5.2. Discussion of Results

We report the results with the best hyperparameter values for each baseline, as shown in Table 2. REEDIT outperforms strong baselines in SSIM, LIPIS, Dir. Similarity, and S-Visual, and is competitive in CLIP Score. VISII optimizes the latent instruction c_T with eight random seeds per sample, and selects the c_T that minimizes the loss term. In contrast, REEDIT does not use repeated generations but still performs well on average, highlighting its computational efficiency, which is crucial for practical editing applications. These findings are further supported by qualitative examples in Fig. 3, where REEDIT demonstrates superior performance across various edit types. Next, we discuss our key observations and results from Fig 3.

Style Transfer. In Rows 1-2, REEDIT successfully captures the stylistic edit from the exemplar pair and applies it to the target image, while completely maintaining the original structure. In both cases, VISII captures the desired style from the exemplar pair, but is unable to maintain the structure of the y image while applying the edit. InstructPix2Pix on the other hand does not sufficiently capture the stylistic information, showing the inability of text alone to sufficiently capture the edit.

Subject/Background Editing. These edits require addition or replacement in the subject or background of the image, while leaving other elements unchanged. REEDIT is able to capture and apply these edits while causing less visual disruption compared to the baselines. In Row 5, only REEDIT is successful at changing the background to an ocean while keeping the horses intact. REEDIT is able to add the subtle raindrops from the exemplar pair in Row 6. Further, only REEDIT is able to capture *all* aspects of the desired edit in Row 4, while other baselines only edit the hat, and not the man's shirt.

Localized Editing. Rows 7-8 show the ability of REEDIT to capture and apply fine-grained edits, such as changing the woman's dress to a suit, or altering her appearance without changing the dressing, while keeping all other elements

Metric	IP2P w/ LLaVA edit text [4]	VISII w/ LLaVA edit text	VISII [34]	**REEDIT** (Ours)
LPIPS [61] (↓)	0.33	<u>0.27</u>	0.29	**0.26**
SSIM [53] (↑)	0.46	**0.51**	<u>0.48</u>	**0.51**
CLIP Score [16] (↑)	29.77	<u>31.62</u>	**31.75**	31.38
Dir. Similarity	0.03	<u>0.04</u>	<u>0.04</u>	**0.05**
S-Visual [34] (↑)	0.22	<u>0.32</u>	**0.39**	**0.39**

Table 2. Quantitative comparison of our framework REEDIT against strong baselines – VISII (and its modifications), and Instruct-pix2pix (IP2P). Reported are the mean of of five different metrics on our dataset. (best scores in **bold**; second best <u>underlined</u>)

Method	Average Inference Time (s)
REEDIT	120s
VISII	540s
VISII w/ Text	550s
IP2P w/ Text	40s

Table 3. Average running time for different methods. Includes all steps in the respective pipelines. REEDIT is more than ∼4 times faster than the most performant baseline - VISII w/ Text. As shown in Sec. 5, IP2P w/text is not performant in this setting.

largely intact. Here, VISII, and VISII w/ Text introduce noise and artifacts into the test image, while InstructPix2Pix is unable to maintain the background. In Row 8, the subtle change to a wooden structure is perfectly captured and applied to *only the required regions* by REEDIT, while other methods completely fail. Additional qualitative results are presented in Appendix A.3.

6. Ablation Analysis

This section comprises of detailed experiments and ablations of our framework, REEDIT. We ablate on the key aspects in our approach - **a. LLaVa Text** (g_{caption}), **b.** Guidance (f, Q, K injection), and **c. Clip Difference** (Δ_{img}). The results are presented in Fig. 4. Additionally, we show the effect of varying our the edit weight (λ) in Fig. 5. We discuss both sets of results in detail below

Impact of LLaVA Text (g_{caption}). Our approach leverages both image and text guidance to sufficiently capture all aspects of the edit from the exemplar pair (x, x_{edit}). There is a synergestic relation between both guidances, filling in the gaps left by the other. This is clearly seen in Row 2, where removing small details like the flowers atop the cactii are missed in the absence of g_{caption}, and in Row 4, where the LLaVA edit text instructs REEDIT to include the sharp teeth of the shark, a subtle cue not captured in its absence.

Impact of Guidance (f, Q, K). Although Δ_{img} and DDIM inversion both provide cues to maintain the structure of the original image, this is enforced much more robustly through feature injection and self attention injection. In the absense of this component, the edited images fail to maintain the structure from the original image, even though stylistic cues are well captured. This is evident from Row 3, where the person is no longer the same, and Rows 4-8, where the structure of the object has been completely destroyed.

Impact of Image Clip Difference (Δ_{img}). This key component of REEDIT captures all types of nuanced details about the edit from the exemplar pair (x, x_{edit}). As we posit in Sec. 3, certain aspects cannot be sufficiently explained through text, which is where this component because increasingly important. The effect is clearly visible across multiple examples. In Row 1, removing Δ_{img} causes the edited image to miss the required style, and instead simply mimics a generic 'caricature'. In Row 2, it is key in capturing the exact required style. The subtle change to a wooden structure in Row 6 is also perfectly capture only in the presence of Δ_{img}. It is easy for the LLaVA generated edit text to miss this detail, while it is easily picked up when analyzing the edit in the latent image space.

Impact of edit weight (λ). The only hyperparameter while performing inference time edits using REEDIT is λ, the edit weight. This acts as a mixing ratio, weighing the contributions of the exemplar pair (x, x_{edit}) and input image y when generating the final output. The effect of this is portrayed in Fig. 5. As we vary the value of λ, we observe a smooth increase in the infuence of the desired edit on the target image. This is a convenient way for a practitioner to exercise control over the editing process. In practice, we find that a value of 0.65 works best across varied types of edits, and this is also the value that yields the best quantitative results.

7. Conclusion

In this paper, we introduce REEDIT, an efficient, optimization-free framework for exemplar-based image editing. We motivate that precise edits cannot be captured by textual modality alone, and propose a novel strategy that leverages the reasoning capabilities of VLMs, and edits in image space to capture the desired user intent using exemplar pairs. Our results demonstrate REEDIT's practical applicability because of its speed and ease of use compared to strong baselines. Our results also position our method as the state of the art both quantitatively and qualitatively. We hope our findings motivate further research in this area.

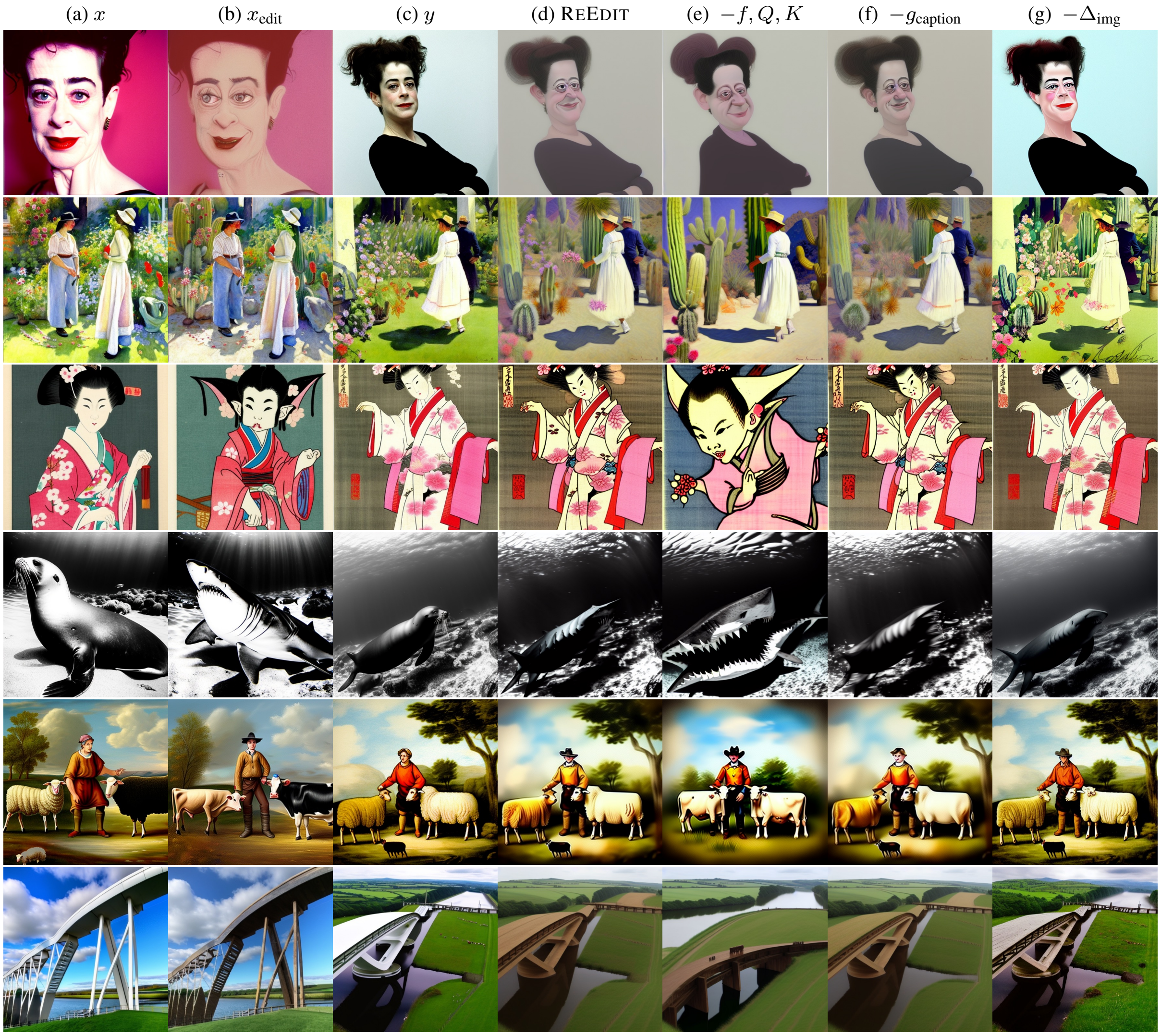

Figure 4. Qualitative results of REEDIT with and without its key components. Clearly, REEDIT outperforms all other variations in terms of adhering faithfully to the edit illustrated in the exemplar without distorting the test image unnecessarily. Use high levels of magnifications to observe subtle edits.

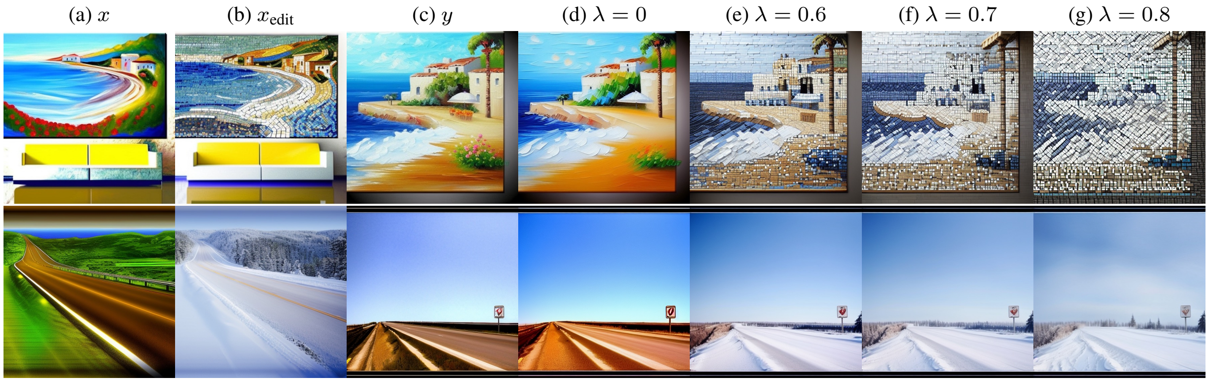

Figure 5. Depiction of the effect of edit weight λ on the edited image. higher values correspond to higher influence from the desired edit.

References

[1] Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. Third time's the charm? image and video editing with stylegan3. In *European Conference on Computer Vision*, pages 204–220. Springer, 2022. 1

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18208–18218, 2022. 2

[3] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017, 2022. 1, 2

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1, 2, 6, 7, 13

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. 2

[6] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021. 2

[7] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 2

[8] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021. 2

[9] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 2

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2

[11] 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 Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 1

[12] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: clip-guided domain adaptation of image generators. corr abs/2108.00946 (2021). *arXiv preprint arXiv:2108.00946*, 2021. 4, 12

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[14] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 2, 4

[15] Aaron Hertzmann, Charles E Jacobs, Nuria Oliver, Brian Curless, and David H Salesin. Image analogies. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 557–570. 2023. 1, 2

[16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 4, 7, 12

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[18] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 6

[19] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. Reversion: Diffusion-based relation inversion from images. *arXiv preprint arXiv:2303.13495*, 2023. 1, 2

[20] Jaeseok Jeong, Junho Kim, Yunjey Choi, Gayoung Lee, and Youngjung Uh. Visual style prompting with swapping self-attention. *arXiv preprint arXiv:2402.12974*, 2024. 1, 2

[21] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1

[22] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 1, 2

[23] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 1, 2

[24] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 2

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2

[26] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017. 2

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2, 3

[28] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 1, 2, 3, 6

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 3

[30] Yunzhe Liu, Rinon Gal, Amit H Bermano, Baoquan Chen, and Daniel Cohen-Or. Self-conditioned generative adversarial networks for image editing. *arXiv preprint arXiv:2202.04040*, 2022. 1

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[32] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 2

[33] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 1, 2

[34] Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via image prompting. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 4, 6, 7, 12

[35] 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. 1, 2

[36] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 1, 2

[37] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[39] 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. 2

[40] 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

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 2

[43] 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. 1, 2

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 12

[45] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 2

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 4

[47] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[48] Adéla Šubrtová, Michal Lukáč, Jan Čech, David Futschik, Eli Shechtman, and Daniel Sỳkora. Diffusion image analogies. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–10, 2023. 1, 2

[49] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 1, 2, 4

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[51] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 2

[52] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 2

[53] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4, 7, 12

[54] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 2

[55] Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, Hideki Koike, et al. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[56] Zhen Yang, Dinggang Gui, Wen Wang, Hao Chen, Bohan Zhuang, and Chunhua Shen. Object-aware inversion and reassembly for image editing. *arXiv preprint arXiv:2310.12149*, 2023. 2

[57] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023. 3, 6

[58] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 2

[59] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 2

[60] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[61] 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. 4, 7, 12

[62] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6027–6037, 2023. 2

[63] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 597–613. Springer, 2016. 2