Mojo: Training-Free Image Editing via Skip Connection Modulation

Peiang Zhao[1] **Han Li**[1] **Ruiyang Jin**[1] **S. Kevin Zhou**[1,2]
[1] University of Science and Technology of China
[2] Institute of Computing Technology, Chinese Academy of Sciences
[1] {pazhao, hanli21, ryjin}@mail.ustc.edu.cn
[2] {s.kevin.zhou}@gmail.com

Abstract

Text-to-image diffusion models have recently garnered significant attention for their ability to create diverse and realistic visual contents. However, adapting these models for real image editing remains challenging. Existing text-guided image editing methods either struggle to achieve effective editing while maintaining the overall image structure, or require extensive fine-tuning, making them impractical for many applications. To address these challenges, we introduce **Mojo**, a novel training-free approach for effective and structure-preserving image editing. Mojo incorporates two innovative techniques: Skip Connection Modulation (SCM) and Cross Image Self-Attention (CISA). SCM leverages the potential of skip connections within the diffusion U-Net. By modulating skip connection features during image editing process, it retains the source image's structure while facilitating successful modifications. CISA further enhances the quality of edited images by improving fine-grained visual details through self-attention transfer. Extensive experiments show that Mojo outperforms existing image editing methods, delivering superior results in versatile image editing scenarios.

1 Introduction

Large-scale text-to-image (T2I) diffusion models, *e.g.*, Stable Diffusion families [1, 2, 3], DALL·E 3 [4], and Imagen [5], have achieved remarkable diversity and realism in image generation, garnering widespread attention. Building on these advanced models, a growing body of research has recently focused on utilizing their capability for text-guided image editing (TIE), which involves transforming a source image I^{src} into an edited image I^{edit} according to a descriptive target prompt P.

However, leveraging a T2I model for editing real images with a text guidance presents significant challenges. Early methods [6, 7, 8] often involve additional training or fine-tuning, which is time-consuming and may restrict their capability for zero-shot generalization. Moreover, some methods require explicit masks to indicate regions that need to be altered, which are costly to obtain or often simply unavailable. These barriers hinder the practical application of these editing techniques in many real-world scenarios.

Subsequent studies have introduced training-free TIE methods. Motivated by image inversion techniques like DDIM inversion [9], these methods establish an *Inversion-Reconstruction* process to extract reference features. These features are then injected into the text-guided *Editing* process to generate edited images while preserving the structure of the source image. For instance, MasaCtrl [10] and FreePromptEditing (FPE) [11] inject self-attention matrices into the editing process. Plug-and-Play (PnP) [12] proposes U-Net backbone spatial features and self-attention overriding for structure preservation. However, these methods often yield unsatisfactory editing results. While self-attention injection excels at preserving fine-grained visual details, the lack of semantic alignment

Preprint. Under review.

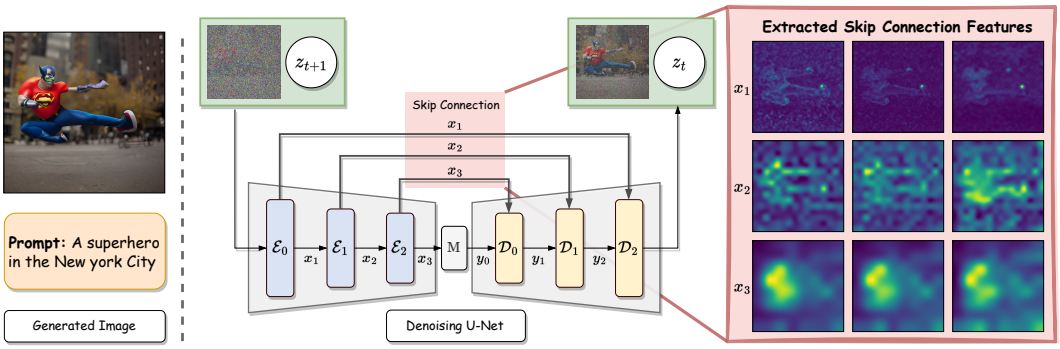

Figure 1: Variance maps across channels of skip connection features within the U-Net for the text prompt "A superhero in New York City" at the *15*-th timestep of the diffusion process. These skip connection features exhibit high variance in regions corresponding to the structure of the generated image.

between the reference and editing self-attention features makes these methods heavily dependent on hyper-parameter selection to achieve satisfactory outcomes. Sub-optimal hyper-parameters frequently lead to over-editing, where the image structure is drastically altered, particularly in images with complex structures or distinct art styles. Conversely, they may also cause negligible image alterations, resulting in editing failures. Injecting backbone spatial features helps alleviate this issue but can result in the loss of visual details in the edited images. This is primarily because these backbone features mainly consist of low-frequency information, as analyzed in [13].

Recently, a noteworthy line of research has investigated the role of **Skip Connections** between the encoder and decoder of the diffusion U-Net. Researchers have shown the effectiveness of tuning skip connections across diverse tasks, such as enhancing training stability [14], augmenting generation quality [13], and facilitating controllable image synthesis [15]. Inspired by these findings, we further explore the potential of skip connections within a pre-trained diffusion U-Net. As shown in Figure 1, we visualize the variance maps of intermediate features within skip connections during the image generation process. These features exhibit elevated variances in regions corresponding to the generated image's structure, indicating a significant connection between skip connection features and the image structural information.

In light of this revelation, we introduce **Mojo**, a training-free method for precise and consistent text-guided image editing. Mojo exploits the skip connections within the diffusion U-Net for effective image editing while preserving the structure of the source image.

Given a source image I^{src} and a target prompt P, we employ DDIM inversion [9] to obtain its corresponding initial latent z_T^{src} for I^{src}. During this inversion process, we extract the intermediate features of skip connections as a reference. Subsequently, we generate the edited image I^{edit} with z_T^{src} and P, wherein we modulate the skip connections within the diffusion U-Net based on the reference features. However, simply replacing the skip connections with reference features imposes excessively strong structural constraints, resulting in editing failure and deterioration in image quality. To address this, we propose a **S**kip **C**onnection **M**odulation (SCM) mechanism. Specifically, SCM conducts a similarity-aware modulation of skip connections, leveraging semantic similarity to align target structure with the reference while ensuring faithful image editing. Since SCM exclusively manipulates the skip connections, it preserves the structure of the source image effectively while not restricting the interaction between the generation process and the conditioning textual prompt P. Nevertheless, we observe that relying solely on SCM for image editing may degrade fine-grained visual details in the edited results. To mitigate this, we further propose a **C**ross **I**mage **S**elf-**A**ttention (CISA) mechanism. CISA bridges the reconstruction and editing process via self-attention transfer, enhancing the method's capability to generate high-quality edited images.

We conduct comprehensive experiments on various benchmarks, comparing our method with existing training-free text-guided image editing approaches. Mojo demonstrates strong performance across various editing scenarios, showcasing improvements both quantitatively and qualitatively over prior methods.

In summary, our contributions are as follows:

- We propose Mojo, an effective method in leveraging Skip Connections within the diffusion U-Net for precise and faithful text-guided image editing. Mojo requires no extra-training and test-time optimization.
- We introduce two novel mechanisms: Skip Connection Modulation (SCM) and Cross Image Self-Attention (CISA). SCM modulates the skip connections during the generation of the edited image, enabling effective editing while preserving the structure of the source image. CISA enhances the generation of fine-grained details in the edited image by linking the reconstruction and editing processes.
- Experiments show that Mojo achieves outstanding performance in text-guided image editing, outperforming prior state-of-the-art approaches both quantitatively and qualitatively.

2 Related Works

2.1 Text-guided Image Editing (TIE)

TIE is a challenging task that aims to manipulate an image according to a descriptive prompt. Early GAN-based TIE methods excel in specific domains like faces but suffer from limited generalization capability. Below, we mainly discuss diffusion-based TIE approaches, which can be broadly categorized into two types: training-based methods and training-free methods.

Training-based Methods. Training-based TIE methods [6, 8, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28] aim to perform ideal image editing by training an entirely new model or fine-tuning existing ones. For instance, InstructPix2Pix [6] learns to follow the editing instructions with large-scale dataset consists of synthetic "input-instruction-goal" triples. Imagic [8] tunes a pre-trained T2I model to estimate an optimal text embedding of the source image and then conducts editing by interpolating between estimated embedding and target text embedding. More recently, MGIE [17] and SmartEdit [29] propose end-to-end joint training schemes to leverage the potential of Multimodal Large Language Models (MLLMs) for text-guided image editing. While these methods yield noteworthy results, they encounter challenges such as computational inefficiency and labor-intensive data collection for training. In contrast, our method focuses on training-free techniques without any additional training process.

Training-free Methods. Another series of methods [10, 11, 12, 30, 31, 32, 33, 34, 35] achieves TIE by designing a concise training-free scheme with pre-trained T2I models. SDEdit [36] perturbs the source image with Gaussian noise and progressively denoises the pertubated image, resulting in an edited image. Subsequent approaches such as Prompt-to-Prompt [30] and DiffEdit [37] explore the implicit grounding capability of T2I models to estimate the target region automatically, improving image structure preservation. Other recent methods like Plug-and-Play (PnP) [12] and FreePromptEditing [11] propose attention control mechanisms to preserve image structure while manipulating visual content. Built upon effective inversion techniques, these attention-based methods formulate a dual-branch framework that is composed of an *Inversion-Reconstruction* branch and an *Editing* branch, and perform cross-branch attention manipulation to achieve sturcture consistency. However, these methods often produce unstable editing results, either over-editing, which incorrectly alters the image structure or object pose, or under-editing, resulting in no significant change. In contrast, our method leverages structural information in skip connections to better control the structure of the edited image, leading to precise and faithful edits.

2.2 Skip Connections within Diffusion U-Net

U-Net [38] is originally introduced for diffusion models in DDPM [39] and serves as the core architecture of most current diffusion models. Recently a growing body of research [13, 15, 40, 41, 42] has focused on the significance of skip connections in the diffusion U-Net. ScaleLong [14] suggests scaling the coefficients of skip connections to enhance training stability. FreeU [13] re-weights the contributions from the skip connections and backbone feature maps to improve generation quality. SCEdit [15] introduces trainable modules into skip connections for more controllable image synthesis. Inspired by these works, we aim to leverage information in skip connections to achieve effective and structure-preserving image editing.

3 Preliminaries

Diffusion Models. Diffusion models sample high-quality images from Gaussian noise. They operate through a forward process that gradually perturbs initial data z_0 with Gaussian noises:

$$z_t = \sqrt{\alpha_t} z_0 + \sqrt{1-\alpha_t}\varepsilon \quad \varepsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{1}$$

A network ε_θ (typically a U-Net) is then trained to predict the added noise, following the objective:

$$\min_\theta E_{z_0,\varepsilon\sim N(0,I),t\sim\text{Uniform}(1,T)} \left\|\varepsilon - \varepsilon_\theta(z_t, t, \mathcal{C})\right\|_2^2, \tag{2}$$

where $\mathcal{C} = \psi(P)$ is the conditioning text embedding. During inference, we can use DDIM sampling to progressively denoise a noise vector z_t:

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \varepsilon_\theta(z_t, t, \mathcal{C}). \tag{3}$$

In this paper, we use a latent diffusion model, *i.e.*, the publicly available Stable Diffusion [1], in which z_0 is the latent encoding of a real image I_0.

DDIM Inversion. Manipulating real images with diffusion models is challenging due to the absence of their corresponding initial latent representations. Therefore, a technique for real image inversion (*i.e.*, inverting z_0 to z_T) is necessary. A straightforward inversion technique for DDIM sampling assumes that the ODE process can be reversed in the limit of small steps:

$$z^*_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} z^*_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \varepsilon_\theta(z^*_t, t). \tag{4}$$

Here, the diffusion process is performed in reverse, transforming the encoding z_0 to z_T, where z_0 is set to be the encoding of the given source image I^{src} that needs to be edited.

U-Net Architecture. The U-Net architecture, depicted in Figure 1, consists of cascaded encoder and decoder blocks, which first encode the input into high-level representations and then decode them to produce an output. Each encoder and decoder block comprises a residual block, a self-attention module, and a cross-attention module.

At timestep t, we denote the intermediate features within the U-Net encoder and decoder with x_i and y_i, respectively. For the i-th encoder block $\mathcal{E}_i$, we have:

$$x_{i+1} = \mathcal{E}_i(x_i), \quad 0 \le i \le N-1. \tag{5}$$

Here, t is omitted for clarity. N denotes the total number of encoder blocks. x_{i+1} is the output of $\mathcal{E}_i$. Notably, x_0 serves the input of the U-Net. At denoising timestep t, we have $x_0 = z_t$.

The encoder and decoder blocks are interconnected with additional skip connections. Specifically, for the j-th decoder block $\mathcal{D}_j$, its input is the concatenation of the skip connection feature and the output from previous decoder block:

$$y_{j+1} = \mathcal{D}_j([x_{N-j}; y_j]), \quad 0 \le j \le N-1, \tag{6}$$

where $[;]$ represents the concatenation operation. y_j denotes the output of $\mathcal{D}_{j-1}$. x_{N-j} represents the skip connection feature, which is also the output of encoder block $\mathcal{E}_{N-j-1}$.

Self-Attention Mechanism. The self-attention modules in each encoder and decoder block compute the pairwise similarities among spatial positions within the latent features through the attention operation:

$$\text{SA}\{Q, K, V\} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V, \tag{7}$$

where the queries Q, keys K and values V are projections of the latent feature. As analyzed in [10, 11, 12, 43], the self-attentions hold significant associations with image structure and appearance.

4 Method

Let I^{src} denote the image to be edited, and P is the target prompt. Our objective is to generate an edited image I^{edit} that aligns with P while preserving the composition of I^{src}. To achieve this, Mojo manipulates skip connections of the diffusion U-Net and conducts cross-image self-attention during the generation process. The overall framework is illustrated in Figure 2. Below, we discuss the crucial components of Mojo.

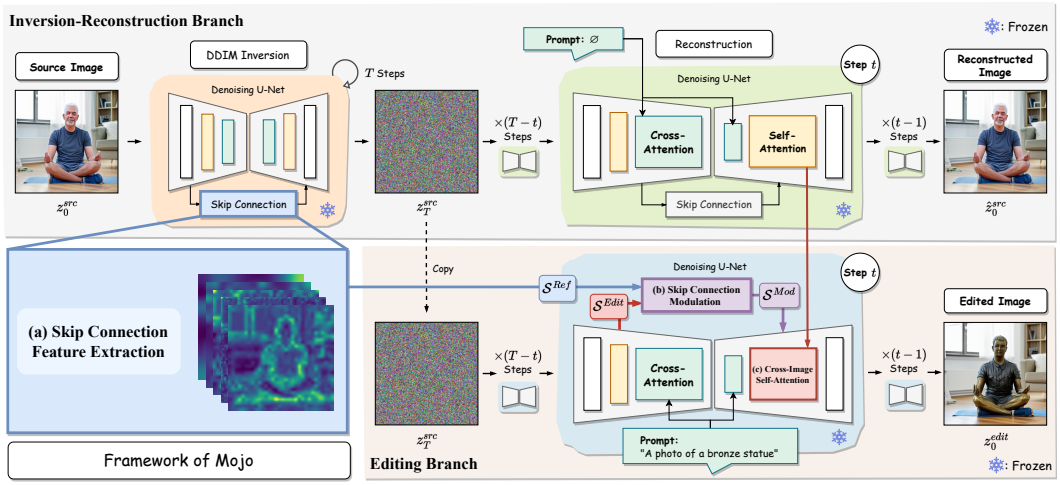

Figure 2: **Framework of Mojo.** We first invert the source image to its corresponding latent encoding via DDIM inversion [9] and extract skip connection features during this process. Subsequently, we simultaneously reconstruct the source image and perform image editing using our proposed Skip Connection Modulation (SCM) and Cross-Image Self-Attention (CISA). Please note that we represent the latent encodings (*e.g.*, z_0^{src}) with their decoded images for clarity.

4.1 Skip Connection Feature Extraction

Starting with a real image I^{src}, we first obtain its corresponding latent encoding z_0^{src}. Subsequenty, z_0^{src} is inverted to the initial noise z_T^{src} via DDIM Inversion and then reconstructed to $\hat{z}_0^{src}$. Throughout the inversion process, we extract skip connection features of the U-Net. Specifically, at timestep t, the set of extracted skip connection features $\mathcal{S}_t^{Inv}$ is defined as:

$$\mathcal{S}_t^{Inv} = \{x_{1,t}^{Inv}, x_{2,t}^{Inv}, ..., x_{N,t}^{Inv}\}, \tag{8}$$

where N denotes the total number of encoder blocks. We then average these features across all timesteps to obtain the reference feature $\mathcal{S}^{Ref}$:

$$\mathcal{S}^{Ref} = \{\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, ..., \bar{\mathbf{x}}_N\}, \quad \bar{\mathbf{x}}_i = \frac{1}{T}\sum_{k=1}^{T} x_{i,k}^{Inv}. \tag{9}$$

where T is the total inversion timestep. According to [13], skip connections introduce high-frequency features with significant variations during the diffusion process. Therefore, averaging features across different timesteps helps to obtain better representations of image structure.

4.2 Skip Connection Modulation (SCM)

Our proposed Skip Connection Modulation (SCM) manipulates skip connection features within the text-conditioned denoising process of z_T^{src} for structure-preserved image editing, as illustrated in Figure 2 (b). For structure preservation, a basic approach would be to directly replace the skip connection features within the denoising process with the reference features $\mathcal{S}^{Ref}$. However, this simplistic strategy imposes excessively strong structural constraints, resulting in editing failures and deterioration in image quality (see Figure 5).

To address this issue, SCM performs a similarity-aware modulation of skip connections to align the target structure with the reference structure while ensuring successful image editing. SCM is applied in the early steps of the editing process ($t \leq \tau_s$). At timestep t, we extract the skip connection features of the editing branch:

$$\mathcal{S}^{Edit} = \{x_1, x_2, ..., x_N\}. \tag{10}$$

For clarity, we omit t in the notation. We then modulate the skip connection features with SCM. The SCM operation $\mathbf{SCM}[;]$ is defined as follows:

$$\mathcal{S}^{Mod} = \mathbf{SCM}[\mathcal{S}^{Ref}; \mathcal{S}^{Edit}], \quad \mathcal{S}^{Mod} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_N\}, \tag{11}$$

where $\mathcal{S}^{Mod}$ represents the set of modulated features and $\mathbf{c}_i$ is the i-th modulated feature. Specifically, SCM evaluates the similarity of features within $\mathcal{S}^{Ref}$ and $\mathcal{S}^{Edit}$ at each spatial location (k, l) and performs skip connection modulation accordingly. For each feature pair (*i.e.*, $\bar{\mathbf{x}}_i$ and x_i) within $\mathcal{S}^{Ref}$ and $\mathcal{S}^{Edit}$, we have:

$$\mathbf{c}_i^{(k,l)} = \begin{cases} x_i^{(k,l)} + \eta(\bar{\mathbf{x}}_i^{(k,l)} - x_i^{(k,l)}) & sim^{(k,l)} > \lambda, \\ x_i^{(k,l)} & otherwise. \end{cases} \tag{12}$$

where $\mathbf{c}_i$ represents the i-th modulated feature, $sim^{(k,l)}$ denotes the similarity between $\mathbf{x}_i^{(k,l)}$ and $x_i^{(k,l)}$, and λ is a predefined threshold. The similarity $sim^{(k,l)}$ is defined as:

$$sim^{(k,l)} = \frac{\bar{\mathbf{x}}_i^{(k,l)} \cdot x_i^{(k,l)}}{\|\bar{\mathbf{x}}_i^{(k,l)}\| \cdot \|x_i^{(k,l)}\|}, \tag{13}$$

where $\|\cdot\|$ denotes l-2 normalization. If the similarity exceeds the threshold, we prioritize the reference features; otherwise, we use features from the editing process.

This similarity assessment helps determine which parts of the image need alteration. For example, if the goal is to replace a foreground object, the object area should undergo changes, while the background should ideally remain unchanged. We have observed that skip connection feature pairs corresponding to the object area show deviations, while those in the background show a high degree of similarity. Therefore, in the object area, we retain the features in $\mathcal{S}^{Edit}$ for successful editing. In areas where changes are unnecessary, we mildly inject the reference feature through a weighted summation to preserve the structure. Further discussions and visualizations are available in Appendix A.

Finally, the original skip connection features $\mathcal{S}^{Edit}$ are replaced with the modulated features $\mathcal{S}^{Mod}$ and fed into the decoder blocks:

$$y_{j+1} = \mathcal{D}_j([\mathbf{c}_{N-j}; y_j]), \quad 0 \leq j \leq N-1. \tag{14}$$

4.3 Cross Image Self-Attention (CISA)

Although SCM significantly improves image editing while maintaining structural consistency, it may hinder the generation of fine-grained visual details due to its modulation of high-frequency features within the skip connections.

Self-attention modules compute the pairwise similarities among spatial positions within the latent features after linearly projecting them into queries and keys. Previous studies [12, 43] highlight the link between these affinities and the generation of subtle visual details and structure. Motivated by this insight, we propose the Cross Image Self-Attention (CISA) mechanism to further enhance visual quality, as illustrated in Figure 2 (c). At timestep t ($t \leq \tau_c$), we integrate the queries and keys from the reconstruction process into self-attention modules of the editing branch as follows:

$$\text{SA}^{Edit}\{Q, K, V\} = \text{Softmax}(\frac{QK^T}{\sqrt{d}})V, \tag{15}$$

where

$$\{Q, K, V\} = \begin{cases} \{Q_{Src}, K_{Src}, V_{Edit}\} & L_1 \leq l \leq L_2, \\ \{Q_{Edit}, K_{Edit}, V_{Edit}\} & otherwise. \end{cases} \tag{16}$$

Here, Q_{Src} and K_{Src} represent the queries and keys from the reconstruction process of the source image, while Q_{Edit} and K_{Edit} are their counterparts within the editing process. L_1 and L_2 denote the self-attention layer indexes within the decoder where the injection operation begins and ends, respectively, which we set to 4 and 11.

Please note that we exclusively apply CISA at decoder self-attention layers with low resolutions and early timesteps. This is because, overdoing the attention injection (*e.g.*, injecting across all layers and denoising steps) would result in an edited image that is almost identical to the source image, leading to editing failure. By restricting the injection to low-resolution self-attention layers, we enhance the generation of fine visual details while preventing unwanted appearance leakage from the source image to the edited one.

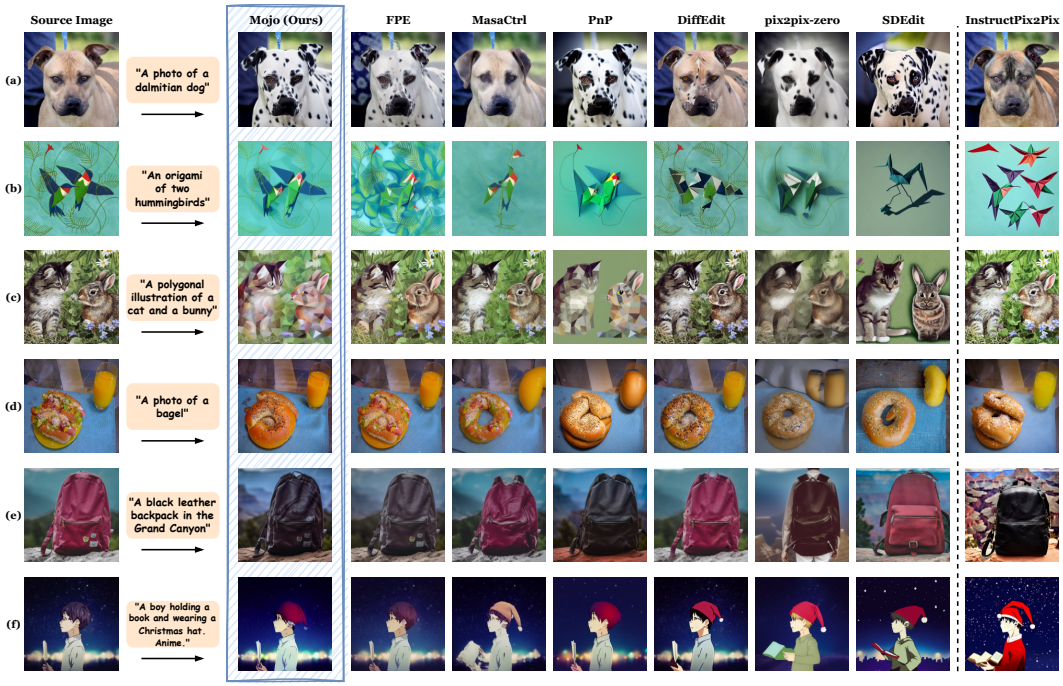

Figure 3: **Visual comparisons.** We compare Mojo with prior text-guided image editing approaches across various editing scenarios, such as object replacement and style transformation. We also introduce training-based InstructPix2Pix [6] for qualitative comparison. Please zoom in for better view.

5 Experiments

5.1 Experimental setups

Datasets. We evaluate Mojo on three benchmarks for real image manipulation. Specifically, we utilize ImageNet-R-TI2I and Wild-TI2I constructed by PnP [12], comprising 150 and 78 image-prompt pairs, respectively. These benchmarks encompass diverse renditions (e.g., paintings, embroidery) of different object classes (e.g., people, animals, landscapes) at various semantic levels (e.g., object-level and scene-level). Additionally, we utilize the ImageNet-Real benchmark introduced by FreePromptEditing [11], which contains 1,092 real images sampled from the ImageNet [44] validation set along with their corresponding editing instructions.

Evaluation metrics. For quantitative analysis and comparison between Mojo and existing training-free image editing apporaches, we employ several metrics. *Clip Score* [45] measures the **semantic alignment** between the generated image and the target prompt, while *DINO-ViT Structure Distance* [46] quantifies the extent of **structure preservation**. *PickScore* [47] is employed to evaluate the **visual quality** of edited images based on learned human preferences. Additionally, we report inference time on one RTX 3090 GPU for efficiency comparison.

Implementation details. We use the official Stable Diffusion V-1.5 [1] as the base model for Mojo and all competing methods. The edited images, with a resolution of 512×512, are generated with 50 denoising steps. For the hyper-parameters, we use $\lambda = 0.8$, $\eta = 0.5$ and $\tau_s = 20$ for Skip Connection Modulation (SCM). τ_c is set to 30 for Cross Image Self-Attention (CISA). Classifier-free guidance [48] is utilized with a fixed guidance scale of 10.5.

5.2 Comparison to Prior/Concurrent Work

In this section, we compare Mojo with state-of-the-art image editing methods, including FreePromptEditing (FPE) [11], MasaCtrl [10], Plug-and-Play (PnP) [12], DiffEdit [37], pix2pix-zero [49], Prompt-to-prompt (P2P) [30], SDEdit [36]. We also include InstructPix2Pix [6], a training-

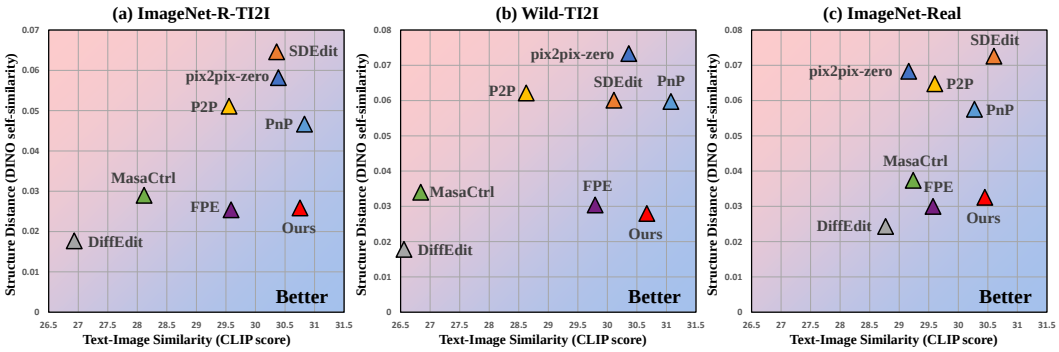

Figure 4: **Quantitative comparisons on Wild-TI2I, ImageNet-R-TI2I and ImageNet-Real benchmarks.** We use the *CLIP score* (higher is better) to quantify the alignment between edited images and target prompts, and the *DINO-ViT self-similarity distance* (lower is better) to measure the preservation of image structure. An ideal image editing method should exhibit high text-image similarity while minimizing structural deviation from the source image.

based method, for qualitative comparison. Detailed discussions on these methods are provided in Appendix D.

Qualitative comparisons. Figure 3 presents visual comparisons of various state-of-the-art image editing methods, demonstrating that Mojo achieves high-quality editing that faithfully aligns with the target prompt while preserving the structural information of the source image. Although FPE and MasaCtrl produce satisfactory results in some cases, they struggle to maintain image structure when transforming the styles of complex images (Figure 3 (b)). DiffEdit excels in object replacement but falls short with texture or style transformations (Figure 3 (c)). SDEdit, pix2pix-zero, and InstructPix2Pix face challenges in preserving the source image's semantic layout (Figure 3 (e)). PnP achieves effective image editing but suffers from degradation of fine-grained visual details (Figure 3 (b)).

Quantitative comparisons. As illustrated in Figure 4, our method surpasses competing methods by achieving excellent structure preservation (low DINO-ViT self-similarity distance) and strong alignment with the target prompt (high CLIP score). In contrast, MasaCtrl and FPE struggle to accurately edit the image according to the target prompt, as evidenced by their lower CLIP scores. Although DiffEdit effectively maintains the image structure by altering only the target region, it falls short in producing faithful edits. SDEdit and PnP achieve successful editing but struggle to retain the original image structure, as indicated by significantly higher DINO-ViT distances. Moreover, Table 1 highlights the superior image quality of Mojo, demonstrated by its higher PickScore compared to competing methods. Although PnP shows comparable performance, its two-stage nature imposes a significant computational burden. In contrast, Mojo is computationally efficient, requiring only 21.55 seconds on an RTX 3090 to manipulate a real image.

Table 1: Comparisons with training-free text-guided image editing methods in terms of **image quality** and **computational efficiency**. PickScore [47] evaluates the quality of edited images based on learned *human preferences*.

Method	**Mojo**	FPE	P2P	PnP	MasaCtrl	DiffEdit	pix2pix-zero	SDEdit
PickScore (↑)	**0.1650**	0.1483	0.1233	0.1411	0.1094	0.0886	0.1319	0.0925
Inference Time (s) (↓)	21.55	21.22	127.98	201.57	21.62	8.95	31.54	**4.23**

5.3 Ablation Study

Qualitative ablations. As observed in Figure 5, generating images directly with target prompt P and z_T^{src} leads to significant structure deviation from the source image. While the proposed CISA helps to some extent in preserving the structure, it still results in erroneous changes. For instance, the generated black eagle lacks a head, and the bear's pose is significantly altered. Solely utilizing the proposed SCM retains the structure of the source image faithfully while ensuring successful editing.

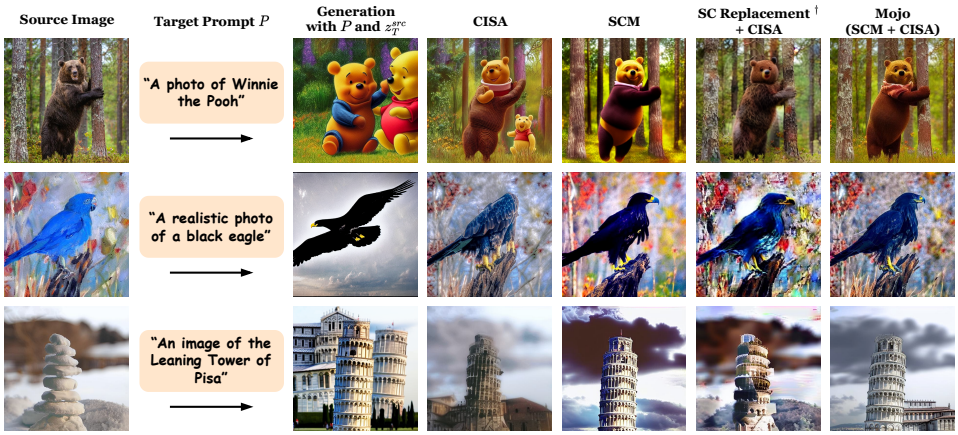

Figure 5: **Visual ablations.** SC Replacement† indicates directly replacing $\mathcal{S}^{Mod}$ with $\mathcal{S}^{Ref}$. Our proposed SCM and CISA collaborate in a complementary manner, achieving an optimal balance between image structure consistency and effective editing.

However, SCM suffers from the deterioration of fine-grained visual features. For example, the black eagle exhibits excessive smoothing and a lack of details. Combining SCM and CISA enables effective and structure-consistent image editing, highlighting the critical role of both components. Additionally, we compare SCM with the naive skip connection modulation strategy discussed in Section 4.2 (*i.e.*, directly replacing $\mathcal{S}^{Mod}$ with $\mathcal{S}^{Ref}$). The results reveal excessively strong structure constraints, leading to editing failure and significant degradation in image quality. This underscores the effectiveness of SCM in achieving high-quality, structure-preserving image editing.

Table 2: **Quantitative ablations.** The results underscores the crucial roles of both SCM and CISA in facilitating faithful image editing. †: *SC Replacement* denote the naive skip connection modulation strategy discussed in Section 4.2.

Method	Wild-TI2I		ImageNet-R-TI2I	
	CLIP Score (↑)	DINO-ViT Distance (↓)	CLIP Score (↑)	DINO-ViT Distance (↓)
SCM	30.7426	0.0684	30.8139	0.0623
CISA	30.4128	0.0458	31.0491	0.0276
SC Replacement† + CISA	27.5693	0.0481	28.2564	0.0394
Mojo (SCM + CISA)	30.6689	0.0279	30.7541	0.0259

Quantitative ablations. Table 2 underscores the crucial roles of both SCM and CISA in facilitating faithful image editing. While SCM preserves the composition of the edited image, it leads to an increase in DiNO-ViT structural distances due to the loss of fine visual details. Conversely, CISA performs well on ImageNetR-TI2I but introduces significant structural changes on Wild-TI2I. This issue arises from the diverse editing scenarios and structurally complex images in Wild-TI2I, which challenge CISA's attention-based mechanism to produce stable results. In contrast, ImageNet-TI2I primarily involves image style transformations (*e.g.*, Figure 3 (b)), where CISA excels as these scenarios require only small changes to the image. Simultaneously utilizing SCM and CISA allows them to complement each other, achieving an optimal balance between image structure consistency and effective editing.

6 Conclusion

In this work, we introduce Mojo, a novel method that harnesses skip connections within the diffusion U-Net for precise and faithful text-guided image editing. We propose Skip Connection Modulation (SCM) to leverage the structural information within the skip connections, facilitating effective and structural-preserving image editing. Additionally, we introduce Cross-Image Self-Attention (CISA) to enhance the generation of fine-grained visual details. Our approach exhibits exceptional performance in both qualitative and quantitative experiments.

References

[1] 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.

[2] 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.

[3] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

[4] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023.

[11] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. *arXiv preprint arXiv:2403.03431*, 2024.

[12] 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.

[13] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. *arXiv preprint arXiv:2309.11497*, 2023.

[14] Zhongzhan Huang, Pan Zhou, Shuicheng Yan, and Liang Lin. Scalelong: Towards more stable training of diffusion model via scaling network long skip connection. *Advances in Neural Information Processing Systems*, 36:70376–70401, 2023.

[15] Zeyinzi Jiang, Chaojie Mao, Yulin Pan, Zhen Han, and Jingfeng Zhang. Scedit: Efficient and controllable image diffusion generation via skip connection editing. *arXiv preprint arXiv:2312.11392*, 2023.

[16] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2(3):5, 2022.

[17] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023.

[18] Shanglin Li, Bohan Zeng, Yutang Feng, Sicheng Gao, Xuhui Liu, Jiaming Liu, Li Lin, Xu Tang, Yao Hu, Jianzhuang Liu, et al. Zone: Zero-shot instruction-guided local editing. *arXiv preprint arXiv:2312.16794*, 2023.

[19] Xue Song, Jiequan Cui, Hanwang Zhang, Jingjing Chen, Richang Hong, and Yu-Gang Jiang. Doubly abductive counterfactual inference for text-based image editing. *arXiv preprint arXiv:2403.02981*, 2024.

[20] Qin Guo and Tianwei Lin. Focus on your instruction: Fine-grained and multi-instruction image editing by attention modulation. *arXiv preprint arXiv:2312.10113*, 2023.

[21] Sherry X Chen, Yaron Vaxman, Elad Ben Baruch, David Asulin, Aviad Moreshet, Kuo-Chin Lien, Misha Sra, and Pradeep Sen. Tino-edit: Timestep and noise optimization for robust diffusion-based image editing. *arXiv preprint arXiv:2404.11120*, 2024.

[22] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. *arXiv preprint arXiv:2311.10089*, 2023.

[23] Hyelin Nam, Gihyun Kwon, Geon Yeong Park, and Jong Chul Ye. Contrastive denoising score for text-guided latent diffusion image editing. *arXiv preprint arXiv:2311.18608*, 2023.

[24] Yuanze Lin, Yi-Wen Chen, Yi-Hsuan Tsai, Lu Jiang, and Ming-Hsuan Yang. Text-driven image editing via learnable regions. *arXiv preprint arXiv:2311.16432*, 2023.

[25] Vidit Goel, Elia Peruzzo, Yifan Jiang, Dejia Xu, Nicu Sebe, Trevor Darrell, Zhangyang Wang, and Humphrey Shi. Pair-diffusion: Object-level image editing with structure-and-appearance paired diffusion models. *arXiv preprint arXiv:2303.17546*, 2023.

[26] Thao Nguyen, Utkarsh Ojha, Yuheng Li, Haotian Liu, and Yong Jae Lee. Edit one for all: Interactive batch image editing. *arXiv preprint arXiv:2401.10219*, 2024.

[27] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023.

[28] Haofeng Liu, Chenshu Xu, Yifei Yang, Lihua Zeng, and Shengfeng He. Drag your noise: Interactive point-based editing via diffusion semantic propagation. *arXiv preprint arXiv:2404.01050*, 2024.

[29] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. *arXiv preprint arXiv:2312.06739*, 2023.

[30] 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.

[31] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with natural language. *arXiv preprint arXiv:2312.04965*, 2023.

[32] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems*, 36, 2024.

[33] Manuel Brack, Felix Friedrich, Katharina Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. *arXiv preprint arXiv:2311.16711*, 2023.

[34] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023.

[35] 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.

[36] 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.

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

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

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

[40] 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.

[41] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

[42] Nithin Gopalakrishnan Nair, Jeya Maria Jose Valanarasu, and Vishal M Patel. Maxfusion: Plug&play multi-modal generation in text-to-image diffusion models. *arXiv preprint arXiv:2404.09977*, 2024.

[43] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. Dreammatcher: Appearance matching self-attention for semantically-consistent text-to-image personalization. *arXiv preprint arXiv:2402.09812*, 2024.

[44] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[45] 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.

[46] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022.

[47] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[49] 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.

[50] 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.