FBSDiff: Plug-and-Play Frequency Band Substitution of Diffusion Features for Highly Controllable Text-Driven Image Translation

Xiang Gao
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
gaoxiang1102@pku.edu.cn

Jiaying Liu*
Wangxuan Institute of Computer Technology, Peking University
Beijing, China
liujiaying@pku.edu.cn

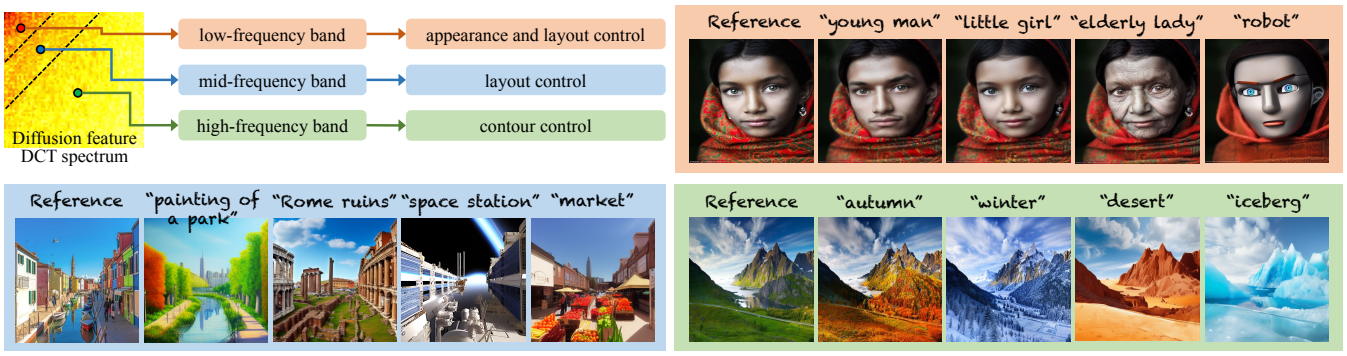

Figure 1: Based on the pre-trained text-to-image diffusion model, FBSDiff enables efficient text-driven image-to-image translation by proposing a plug-and-play reference image guidance mechanism. It allows flexible control over different guiding factors (e.g., image appearance, image layout, image contours) of the reference image to the T2I generated image, simply by dynamically substituting different types of DCT frequency bands during the reverse sampling process of the diffusion model.

Abstract

Large-scale text-to-image diffusion models have been a revolutionary milestone in the evolution of generative AI, allowing wonderful image generation with natural-language text prompt. However, the issue of lacking controllability of such models restricts their practical applicability for real-life content creation. Thus, attention has been focused on leveraging a reference image to control text-to-image synthesis, which is also regarded as manipulating (or editing) a reference image as per a text prompt, namely, text-driven image-to-image translation. This paper contributes a novel, concise, and efficient approach that adapts pre-trained large-scale text-to-image (T2I) diffusion model to the image-to-image (I2I) paradigm in a plug-and-play manner, realizing high-quality and versatile text-driven I2I translation without model training, fine-tuning, or online optimization process. To guide T2I generation with a reference image, we propose to decompose diverse guiding factors with different frequency bands of diffusion features in the DCT spectral space, and accordingly devise a novel frequency band substitution layer which realizes dynamic control of the reference image to the T2I generation result in a plug-and-play manner. We demonstrate that our method allows flexible control over both guiding factor and guiding intensity of the reference image simply by tuning the type and bandwidth of the substituted frequency band, respectively. Extensive qualitative and quantitative experiments verify superiority of our approach over related methods in I2I translation visual quality, versatility, and controllability. Our project is publicly available at: https://xianggao1102.github.io/FBSDiff_webpage/.

CCS Concepts

• **Computing methodologies** → **Image processing**; *Image representations*; Computational photography.

Keywords

Image-to-image translation, Image manipulation, Diffusion model

ACM Reference Format:
Xiang Gao and Jiaying Liu. 2024. FBSDiff: Plug-and-Play Frequency Band Substitution of Diffusion Features for Highly Controllable Text-Driven Image Translation. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3664647.3681380

*Corresponding author

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3681380

1 Introduction

Text-driven I2I translation is an appealing computer vision problem that aims to translate a reference image as per a text prompt. It extends text-to-image (T2I) synthesis to more controllability by controlling T2I generation result with a reference image. Since the

advent of CLIP [28] bridging vision and language through large-scale contrastive pre-training, attempts have been made to instruct image manipulation with text by combining CLIP with generative models. VQGAN-CLIP [6] pioneers text-driven image translation by optimizing VQGAN [9] latent image embedding with CLIP text-image similarity loss. DiffusionCLIP [16] fine-tunes diffusion model [12] under the same CLIP loss to manipulate an image with a text. DiffuseIT [17] combines VIT-based structure loss [38] and CLIP-based semantic loss to guide diffusion model's reverse sampling process via manifold constrained gradient [5], synthesizing translated image that complies with the target text while maintaining the structure of the reference image. However, these methods are not competitive in generation visual quality due to the limited model capacity of backbone generative model as well as the unstability caused by online fine-tuning or optimization process.

To promote image translation visual quality, efforts have been made to train large models on massive data. InstructPix2Pix [2] employs GPT-3 [3] and Stable Diffusion [30] to synthesize huge amounts of paired training data, based on which trains a supervised text-driven I2I mapping for general image manipulation task. Design Booster [36] trains a latent diffusion model [30] conditioned on both text embedding and image embedding, realizing layout-preserved text-driven I2I translation. Nevertheless, these methods are computationally intensive in training large models from scratch and less efficient in collecting immense training data.

To circumvent formidable training costs, research has been focused on leveraging off-the-shelf large-scale T2I diffusion models for text-driven I2I translation. This type of methods further divide into fine-tuning-based methods and inversion-based methods.

The former type of fine-tuning-based methods represented by SINE [43] and Imagic [15] fine-tune the pre-trained T2I diffusion model to reconstruct an input reference image before manipulating it with a target text. These methods require separate fine-tuning of the entire model for each text prompt, which is less efficient and prone to underfitting or overfitting to the reference image.

The latter type of inversion-based methods invert reference image into diffusion model's Gaussian noise space and then generate the translated image via the reverse sampling process guided by the target text. A pivotal challenge of this pipeline is that the sampling trajectory may severely deviate from the inversion trajectory due to the error accumulation caused by the classifier-free guidance technique [13], which severely impairs the correlation between the reference image and the translated image. To remedy this issue, Null-text Inversion [22] optimizes the unconditional null-text embedding to calibrate the sampling trajectory step by step. Prompt Tuning Inversion [8] proposes to minimize trajectory divergence with an optimization to encode the reference image into a learnable prompt embedding. Similarly, StyleDiffusion [18] opts to optimize the "value" embedding of the cross-attention layer as the visual encoding of the reference image. Pix2Pix-zero [25] penalizes trajectory deviation by matching cross-attention maps between the two trajectories with least-square loss. These methods apply per-step online optimization to calibrate the whole sampling trajectory, introducing additional computational cost and time overhead. Moreover, most of these methods adopt the cross-attention control technique introduced in Prompt-to-Prompt [11] for image structure preservation. This makes them rely on a paired source text of the reference image, which is not flexible or even available in most cases. Plug-and-Play (PAP) [39] leverages feature maps and self-attention maps extracted from internal layers of the denoising U-Net to maintain image structure, realizing optimization-free text-driven I2I translation. However, the algorithm is sensitive to specific layer selection, the feature extraction process is also time-consuming.

In this paper, we propose a concise and efficient approach termed FBSDiff, realizing plug-and-play and highly controllable text-driven I2I translation from a frequency-domain perspective. To guide T2I generation with a reference image, a key missing ingredient of existing methods is the mechanism to control the guiding factor (e.g., image appearance, layout, contours) and guiding intensity of the reference image. Since different image guiding factors are difficult to isolate in the spatial domain, we consider decomposing them in the frequency domain by modeling them with different frequency bands of diffusion features in the Discrete Cosine Transform (DCT) spectral space. Based on this motivation, we propose an inversion-based text-driven I2I translation framework featured with a novel frequency band substitution mechanism, which efficiently enables reference image guidance of the T2I generation by dynamically substituting a certain DCT frequency band of diffusion features with the corresponding counterpart of the reference image along the reverse sampling process. As displayed in Fig. 1, T2I generation with appearance and layout control, pure layout control, and contour control of the reference image can be respectively realized by transplanting low-frequency band, mid-frequency band, and high-frequency band between diffusion features, allowing versatile and highly controllable text-driven I2I translation.

The strengths of our method are fourfold: (I) plug-and-play efficiency: it extends pre-trained T2I diffusion model to the realm of I2I in a plug-and-play manner; (II) conciseness: it dispenses with the need for the paired source text of the reference image as well as cumbersome attention modulation process as compared with existing advanced methods, all while achieving leading I2I translation performance; (III) model generalizability: it transplants frequency band of diffusion features along the reverse sampling trajectory, requiring no access to any internal features of the denoising network, and thus decouples with the specific diffusion model backbone architecture as compared with existing methods; (IV) controllability: it allows flexible control over the guiding factor and guiding intensity of the reference image simply by tuning the type and bandwidth of the substituted frequency band. To summarize, we make the following key contributions:

- We provide new insights about controllable diffusion sampling process from a novel frequency-domain perspective.
- We propose a novel frequency band substitution technique, realizing plug-and-play text-driven I2I translation without any model training, model fine-tuning, and online optimization process.
- We contribute a concise and efficient text-driven I2I framework that is free from source text and cumbersome attention modulation operations, highly controllable in both guiding factor and guiding intensity of the reference image, and invariant to the architecture of the used diffusion model backbone, all while achieving superior I2I translation performance compared with existing advanced methods.

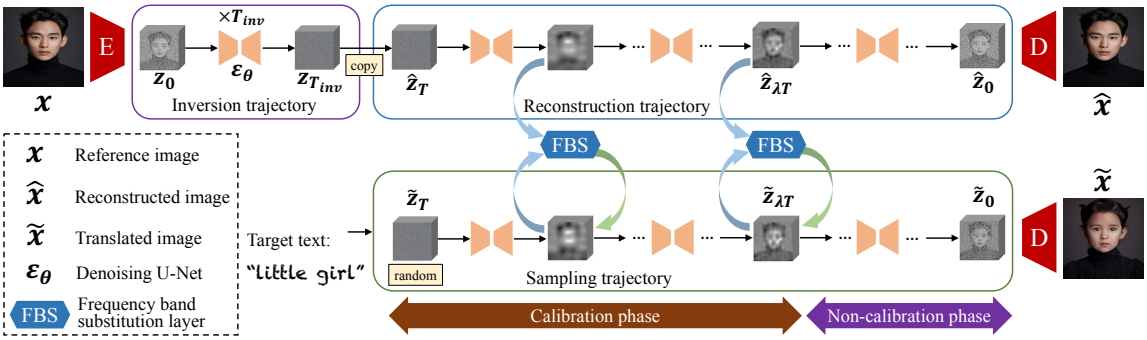

Figure 2: Overview of FBSDiff. Based on the pre-trained latent diffusion model (LDM), FBSDiff starts with an inversion trajectory that inverts reference image into the LDM Gaussian noise space, then a reconstruction trajectory is applied to reconstruct the reference image from the inverted Gaussian noise, providing intermediate denoising results as pivotal guidance features. The guidance features are leveraged to guide the text-driven sampling trajectory of the LDM to exert reference image control, which is realized by inserting our proposed frequency band substitution layer in between the reconstruction and sampling trajectory.

2 Related Work

2.1 Diffusion Model

Since the advent of DDPM [12], diffusion model has soon dominated the family of generative models [7]. DDIM [35] and its variants [20, 20] accelerate diffusion model sampling process to tens of times, promoting its practicability dramatically. Palette [31] extends diffusion model from unconditional image generation to conditional paradigm, opening the door of diffusion-based image-to-image translation. With the advancement of multimodal technology, large-scale T2I diffusion models [24, 29, 32] are proposed to generate high-resolution images with open-domain text prompts, bringing content creation to an unprecedented level. Latent Diffusion Model (LDM) [30] transfers T2I diffusion model from high-dimension pixel space to low-dimensional feature space, reducing the computational overhead significantly. To improve image generation controllability, ControlNet [41] and T2i-adapter [23] add spatial control to T2I diffusion models by training a control module of the denoising U-Net conditioned on certain image priors (e.g., canny edges, depth maps, human key points, etc.). SDXL [27] and DiTs [26] propose Transformer-based denoising network, improving T2I diffusion model to larger capacity. Currently, diffusion model has been making rapid progress in various vision applications such as super-resolution [33], inpainting [21], colorization [19], segmentation [37], 3D reconstruction [1], etc.

2.2 Computer Vision in Frequency Perspective

Research reveals that performance of deep neural networks can be boosted from frequency domain perspective. For example, Ghosh et al. [10] introduce DCT to CNN for image classification, accelerating network convergence speed. Xie et al. [40] propose a frequency-aware dynamic network for lightweight image super-resolution. Cai et al. [4] impose Fourier frequency spectrum consistency to image translation tasks, achieving better identity preservation. FreeU [34] improves T2I generation quality by selectively enhancing or depressing different frequency components of diffusion features inside the denoising U-Net model. In this work, we solve text-driven I2I problem through dynamic DCT frequency band substitution.

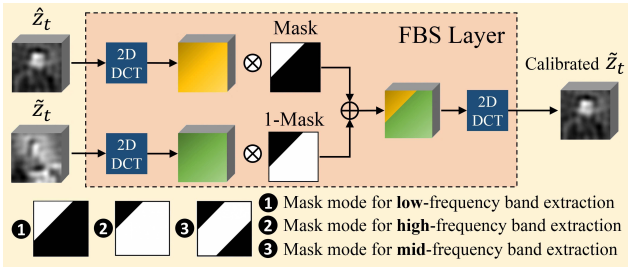

Figure 3: Illustration of the frequency band substitution (FBS) layer. The FBS layer takes in two diffusion features and substitutes a certain frequency band of one diffusion feature with the corresponding frequency band of the other feature in 2D DCT domain. Better viewed with zoom-in.

3 Method

3.1 Overall Architecture

Established on the pre-trained Latent Diffusion Model (LDM), FBSDiff adapts it from T2I generation to text-driven I2I translation with our proposed dynamic frequency band substitution, which efficiently realizes flexible control over both guiding factor and guiding intensity of the reference image to the T2I generated image.

As Fig. 2 shows, FBSDiff comprises three diffusion trajectories: (i) inversion trajectory ($z_0 \rightarrow z_{T_{inv}}$); (ii) reconstruction trajectory ($z_{T_{inv}} = \hat{z}_T \rightarrow \hat{z}_0 \approx z_0$); (iii) sampling trajectory ($\tilde{z}_T \rightarrow \tilde{z}_0$). Starting from the initial feature $z_0 = E(x)$ extracted from the reference image x by the LDM encoder E, a T_{inv}-step DDIM inversion is employed to project z_0 into the Gaussian noise latent space conditioned on the null-text embedding $v_\emptyset$, based on the assumption that the ODE process can be reversed in the limit of small steps:

$$z_{t+1} = \sqrt{\bar{\alpha}_{t+1}} f_\theta(z_t, t, v_\emptyset) + \sqrt{1-\bar{\alpha}_{t+1}} \epsilon_\theta(z_t, t, v_\emptyset), \quad (1)$$

$$f_\theta(z_t, t, v_\emptyset) = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, t, v_\emptyset)}{\sqrt{\bar{\alpha}_t}}, \quad (2)$$

where $\{\bar{\alpha}_t\}$ are schedule parameters that follows the same setting as DDPM [12], ϵ_θ is the denoising U-Net of the pre-trained LDM. The

Gaussian noise $z_{T_{inv}}$ obtained after the T_{inv}-step DDIM inversion is directly used as the initial noise feature of the subsequent reconstruction trajectory, which is a T-step DDIM sampling process that reconstructs $\hat{z}_0 \approx z_0$ from the inverted noise feature $\hat{z}_T = z_{T_{inv}}$:

$$\hat{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} f_\theta(\hat{z}_t, t, v_\emptyset) + \sqrt{1-\bar{\alpha}_{t-1}} \epsilon_\theta(\hat{z}_t, t, v_\emptyset), \quad (3)$$

in which $f_\theta(\hat{z}_t, t, v_\emptyset)$ follows the same form as Eq. 2. The length of the reconstruction trajectory could be much smaller than that of the inversion trajectory (i.e., $T \ll T_{inv}$) to save inference time. The reconstruction trajectory is conditioned on the same null-text embedding $v_\emptyset$ to ensure feature reconstructability (i.e., $\hat{z}_0 \approx z_0$).

Meanwhile, an equal-length sampling trajectory is applied in parallel with the reconstruction trajectory for T2I synthesis. The sampling trajectory is also a T-step DDIM sampling process that progressively denoises a randomly initialized Gaussian noise feature $\tilde{x}_T \sim \mathcal{N}(0, I)$ into $\tilde{x}_0$ conditioned on the text embedding v of the target text prompt. To amplify the effect of text guidance, we employ classifier-free guidance technique [13] by interpolating conditional (target text) and unconditional (null text) noise prediction at each time step with a guidance scale ω during sampling:

$$\tilde{z}_{t-1} = \sqrt{\bar{\alpha}_{t-1}} f_\theta(\tilde{z}_t, t, v, v_\emptyset) + \sqrt{1-\bar{\alpha}_{t-1}} \epsilon_\theta(\tilde{z}_t, t, v, v_\emptyset), \quad (4)$$

$$f_\theta(\tilde{z}_t, t, v, v_\emptyset) = \frac{\tilde{z}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\tilde{z}_t, t, v, v_\emptyset)}{\sqrt{\bar{\alpha}_t}}, \quad (5)$$

$$\epsilon_\theta(\tilde{z}_t, t, v, v_\emptyset) = \omega \cdot \epsilon_\theta(\tilde{z}_t, t, v) + (1-\omega) \cdot \epsilon_\theta(\tilde{z}_t, t, v_\emptyset). \quad (6)$$

Due to the inherent property of DDIM inversion and DDIM sampling, the reconstruction trajectory forms a deterministic denoising mapping towards the reference image, during which the intermediate denoising results $\{\hat{z}_t\}$ can function as pivotal guidance features to calibrate the corresponding counterparts $\{\tilde{z}_t\}$ along the sampling trajectory. Thus, correlation between the reference image and the generated image can be established to allow for text-driven I2I translation. Specifically, we implement feature calibration by inserting a plug-and-play frequency band substitution (FBS) layer in between the reconstruction trajectory and the sampling trajectory. FBS layer dynamically substitutes a certain frequency band of $\tilde{z}_t$ with the corresponding frequency band of $\hat{z}_t$ along the reverse sampling process, which efficiently imposes guidance of the reference image to the T2I generation result.

To improve I2I translation visual quality, we partition the sampling process into a calibration phase and a non-calibration phase, separated by the time step λT. In the former calibration phase ($\tilde{z}_T \rightarrow \tilde{z}_{\lambda T}$), dynamic frequency band substitution is applied at each time step for smooth calibration of the sampling trajectory; in the latter non-calibration phase ($\tilde{z}_{\lambda T-1} \rightarrow \tilde{z}_0$), we remove FBS layer to avoid over-constrained sampling result, fully unleashing the generative power of the pre-trained T2I model to improve image generation quality. Here λ denotes the ratio of the length of the non-calibration phase to that of the entire sampling trajectory.

At last, the final result $\tilde{z}_0$ of the sampling trajectory is decoded back to the image space via the LDM decoder D, producing the final translated image $\tilde{x}$, i.e., $\tilde{x} = D(\tilde{z}_0)$.

3.2 Frequency Band Substitution Layer

As Fig. 3 illustrates, the FBS layer takes in a pair of diffusion features $\hat{z}_t$ and $\tilde{z}_t$, converts them from the spatial domain into the frequency domain via 2D DCT, then transplants a certain frequency band in the DCT spectrum of $\hat{z}_t$ to the same location in the DCT spectrum of $\tilde{z}_t$. Finally, 2D IDCT is applied to transform the fused DCT spectrum of $\tilde{z}_t$ back to the spatial domain as the final calibrated feature.

Algorithm 1 Complete algorithm of FBSDiff

Input: the reference image x and the target text.
Output: the translated image $\tilde{x}$.

```
Extract the initial latent feature z_0 = E(x).
for t = 0 to T_inv − 1 do
    compute z_{t+1} from z_t via Eq. 1;
end for{DDIM inversion}
Initialize ẑ_T = z_{T_inv}, z̃_T ~ N(0, I).
for t = T to λT + 1 do
    compute ẑ_{t−1} from ẑ_t via Eq. 3;
    compute z̃_{t−1} from z̃_t via Eq. 4;
    substitute a certain frequency band of z̃_{t−1} with the corresponding counterpart of ẑ_{t−1} via Eq. 7;
end for{DDIM sampling in the calibration phase}
for t = λT to 1 do
    compute z̃_{t−1} from z̃ via Eq. 4;
end for{DDIM sampling in the non-calibration phase}
Obtain z̃_0 and the final translated image x̃ = D(z̃_0).
```

In 2D DCT spectrum, elements with smaller coordinates (nearer to the top-left origin) encode lower-frequency information while larger-coordinate elements correspond to higher-frequency components. We use the sum of 2D coordinates in DCT spectrum as threshold to extract DCT frequency bands of different types and bandwidths via binary masking. Specifically, we design three types of binary masks which are respectively termed the low-pass mask ($Mask_{lp}$), high-pass mask ($Mask_{hp}$), and mid-pass mask ($Mask_{mp}$):

$$\begin{cases} Mask_{lp}(x,y) = 1 \ \ if \ \ x+y \le th_{lp} \ \ else \ \ 0, \\ Mask_{hp}(x,y) = 1 \ \ if \ \ x+y > th_{hp} \ \ else \ \ 0, \\ Mask_{mp}(x,y) = 1 \ \ if \ \ th_{mp1} < x+y \le th_{mp2} \ \ else \ \ 0, \end{cases}$$

where th_{lp} is the threshold of the low-pass filtering; th_{hp} is the threshold of the high-pass filtering; th_{mp1} and th_{mp2} are respectively the lower and upper bound of the mid-pass filtering. Given a binary mask $Mask_* \in \{Mask_{lp}, Mask_{hp}, Mask_{mp}\}$, the frequency band substitution in the FBS layer can be formulated as:

$$\tilde{z}_t = IDCT(DCT(\hat{z}_t) \cdot Mask_* + DCT(\tilde{z}_t) \cdot (1 - Mask_*)), \quad (7)$$

where DCT and $IDCT$ refer to 2D DCT and 2D IDCT transformations. The use of low-pass mask $Mask_{lp}$, high-pass mask $Mask_{hp}$, and mid-pass mask $Mask_{mp}$ respectively corresponds to the extraction and substitution of the low-frequency band, high-frequency band, and mid-frequency band. They control different guiding factors of the reference image to the T2I generation result:
Low-frequency band substitution enables low-frequency information guidance of the reference image x, realizing appearance (e.g., color, luminance) and layout control over the generated $\tilde{x}$;
High-frequency band substitution enables high-frequency information guidance of x, realizing image contour control over $\tilde{x}$;
Mid-frequency band substitution enables mid-frequency information guidance of x, realizing image layout control over $\tilde{x}$.

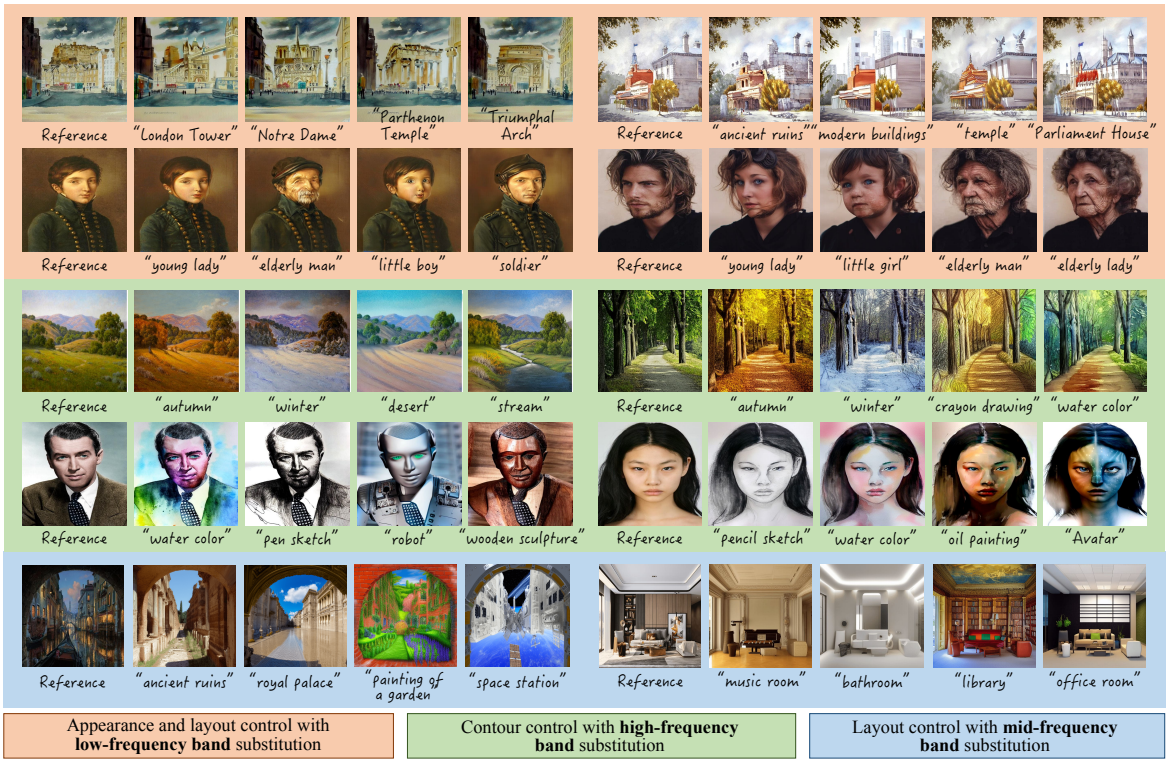

Figure 4: Qualitative results of our method with different types of frequency band substitution. The generated image is controlled by the reference image in terms of image appearance and layout with low-FBS; in terms of image contours with high-FBS; and in terms of pure image layout with mid-FBS. Better viewed with zoom-in.

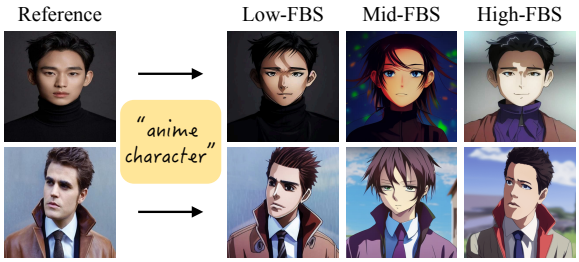

Figure 5: Comparison among different guiding factors achieved by low-FBS, mid-FBS, and high-FBS. The T2I result maintains the appearance and layout of the reference image with low-FBS; preserves image contours with high-FBS; and inherits pure image layout with mid-FBS.

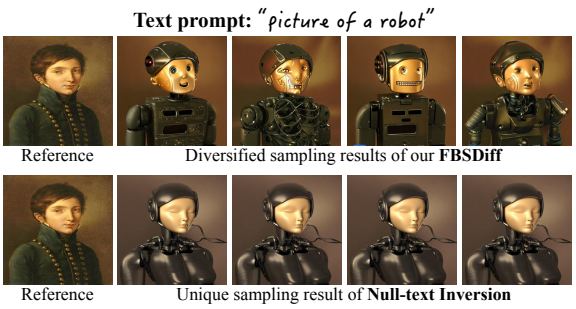

Figure 6: Our method enables diverse sampling results for fixed reference image and text prompt, as contrasted with Null-text Inversion that produces unique sampling result.

The DCT masking type and the corresponding thresholds used in the FBS layer are hyper-parameters of our method, which could be flexibly modulated to enable control over diverse guiding factors and continuous guiding intensity of the reference image.

3.3 Implementation Details

We use the pre-trained Stable Diffusion v1.5 as the backbone diffusion model and set the classifier-free guidance scale $\omega = 7.5$. We use 1000-step DDIM inversion to ensure high-quality reconstruction, i.e., T_{inv}=1000, and use 50-step DDIM sampling for both the reconstruction and sampling trajectory, i.e., T=50. Along the sampling trajectory, we allocate 55% time steps to the calibration phase and the remaining 45% steps for the non-calibration phase, i.e., λ=0.45. For the default DCT masking thresholds used in the FBS layer, we set th_{lp}=80 for low-frequency-band substitution (low-FBS); th_{hp}=5 for high-frequency-band substitution (high-FBS); th_{mp1}=5, th_{mp2}=80 for mid-frequency-band substitution (mid-FBS). The complete algorithm of FBSDiff is presented in Alg. 1.

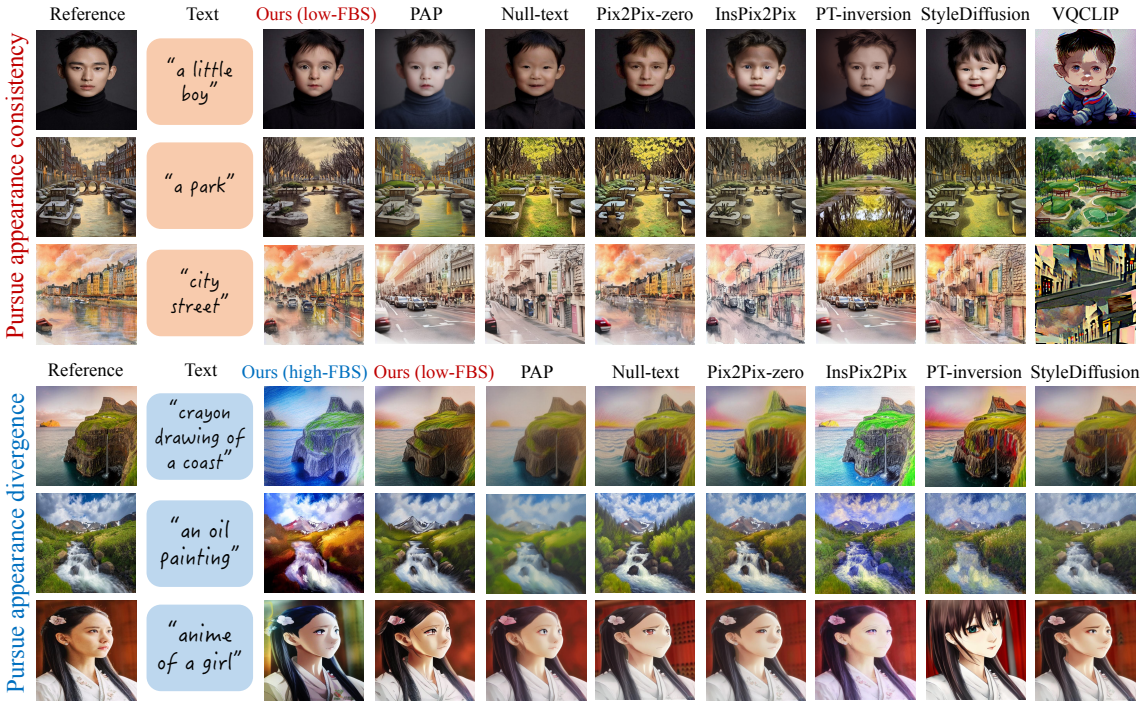

Figure 7: Qualitative method comparisons. Our FBSDiff with low-FBS is more adept at appearance preservation than related methods, which better suits to I2I task pursuing appearance consistency between the reference image and the generated image (top panel). Conversely, our method with high-FBS remarkably facilitates I2I appearance change compared with related methods, which better suits to I2I task pursuing appearance divergence (bottom panel). Better viewed with zoom-in.

Figure 8: Comparison between results with low-FBS and without FBS. Better viewd with zoom-in.

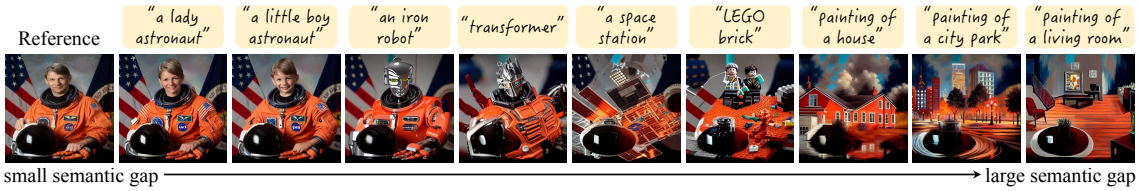

Figure 9: Adaptation to varying semantic gap between the reference image and target text. Better viewed with zoom-in.

4 Experiments

4.1 Qualitative Results

Example text-driven I2I translation results of our method are shown in Fig. 4. The low-FBS produces translated images that inherit the appearance and layout of the reference image. For high-FBS, the generated image is aligned with the reference image in high-frequency contours while the low-frequency appearance is not restricted. Results of mid-FBS only maintain the overall image layout of the reference image, since lower-frequency appearance and the higher-frequency contour information are filtered out in the DCT domain. For all three modes of frequency band substitution, results exhibit both high visual quality and high text fidelity. The control over different guiding factors is more clearly demonstrated in Fig. 5.

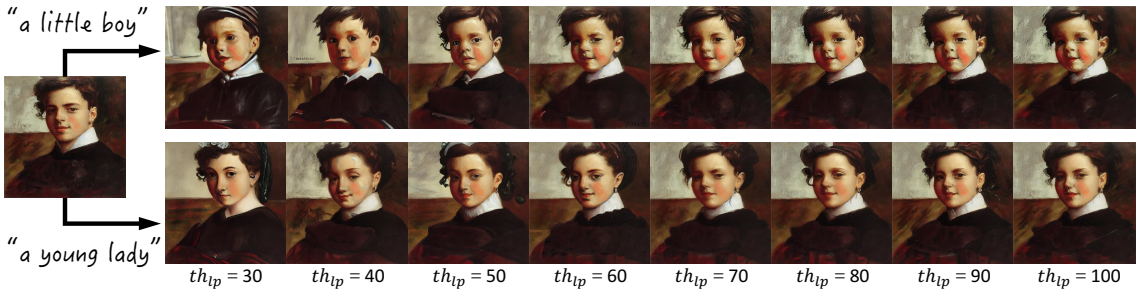

Figure 10: Appearance and layout guiding intensity control realized by varying th_{lp} in low-FBS. Better viewed with zoom-in.

Figure 11: Contour guiding intensity control realized by varying th_{mp2} in mid-FBS. Better viewed with zoom-in.

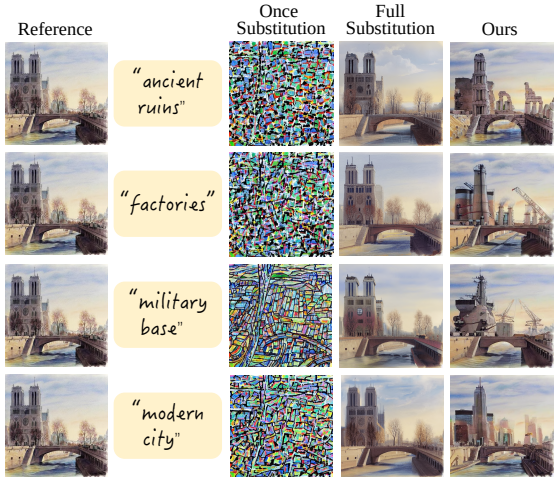

Figure 12: Ablation study w.r.t. different manners of frequency band substitution. Better viewed with zoom-in.

We qualitatively compare our method with SOTA text-driven I2I translation methods including Plug-and-Play (PAP) [39], Null-text Inversion (Null-text) [22], Pix2Pix-zero [25], InstructPix2Pix (InsPix2Pix) [2], Prompt Tuning Inversion (PT-inversion) [8], StyleDiffusion [18], and VQGAN-CLIP (VQCLIP) [6], results are displayed in Fig. 7. The top panel of Fig. 7 shows that our method with low-FBS achieves better I2I appearance consistency than related approaches, and is thus better suited to image creation scenario which favors inheriting the appearance and style from an existing image. The bottom panel of Fig. 7 shows that existing SOTA text-driven I2I methods struggle at producing I2I results with large appearance change from the reference images, while our method with high-FBS excels in generating I2I results with significantly different appearance, and is thus more suitable to image creation scenario where appearance divergence is pursued.

An advantage of our approach over related methods is sampling diversity. As displayed in Fig. 6, our FBSDiff can produce diverse text-guided I2I results by randomly sampling $\tilde{x}_T$ from isotropic Gaussian distribution, while other inversion-based methods [8, 18, 22, 25, 39] lack such sampling diversity due to directly initializing $\tilde{x}_T$ with the inverted feature embedding of the reference image.

The importance of FBS for reference image control is clearly shown in Fig. 8, from which we see that low-FBS establishes I2I appearance and layout correlations, while removing FBS leads to results without any correlation to the reference images. Moreover, as Fig. 9 displays, our method robustly adapts to varying degrees of semantic gap between the reference image and the target text prompt. The translated image of our method can still comply with the target text accurately with satisfying visual quality even in the case of very large image-text semantic discrepancy.

Besides, our method also allows continuous control over the guiding intensity of the reference image simply by modulating the bandwidth of the substituted frequency band. Results displayed in Fig. 10 demonstrate the image appearance and layout guiding intensity control of our method by adjusting the low-pass filtering threshold th_{lp} in the mode of low-FBS. Enlarging the value of th_{lp} widens the bandwidth of the transplanted low-frequency band and thus increases the amount of guiding information of the reference image, leading to the translated image with more resemblance to the reference image. Conversely, lowering the value of th_{lp} narrows the bandwidth of the substituted frequency band, which reduces the amount of guiding information and thus brings more variations to the translated result as compared with the reference image.

Likewise, results in Fig. 11 demonstrate image contour guiding intensity control of our method by adjusting the upper bound threshold th_{mp2} in the mode of mid-FBS. Increasing the value of th_{mp2}

Table 1: Quantitative evaluations of text-driven I2I translation methods.

Emphasis	Pursuing image appearance consistency					Pursuing image appearance divergence			
Metrics / Methods	Structure Similarity(↑)	LPIPS(↓)	AdaIN Style Loss(↓)	CLIP Similarity(↑)	Aesthetic Score(↑)	Structure Similarity(↑)	AdaIN Style Loss(↑)	CLIP Similarity(↑)	Aesthetic Score(↑)
PAP [39]	0.954	0.272	20.440	0.287	6.590	0.956	28.337	0.279	6.458
Null-text [22]	0.948	0.247	17.546	0.276	6.505	0.952	22.545	0.270	6.402
Pix2Pix-zero [25]	0.951	0.243	16.875	0.262	6.484	0.953	21.240	0.258	6.344
InsPix2Pix [2]	0.958	0.266	23.373	0.258	6.269	0.965	30.804	0.264	6.196
PT-inversion [8]	0.947	0.248	21.667	0.271	6.481	0.948	24.367	0.267	6.285
StyleDiffusion [18]	0.944	0.251	22.484	0.267	6.477	0.947	25.166	0.260	6.267
FBSDiff (ours)	0.962	0.241	15.452	0.285	6.583	0.964	33.875	0.281	6.463

The red font indicates the top-ranked value and the blue font indicates the second-ranked value.

leads to more high-frequency components of the reference image included into the transplanted frequency band, and thus results in stronger I2I contour consistency. On the contrary, decreasing the value of th_{mp2} shrinks the transplanted high-frequency guiding information and thus leads to weaker image contour consistency.

4.2 Ablation Study

We also explore other designs of FBS, including substituting the frequency band only once at λT time step rather than along the whole calibration phase (which we denote as **Once Substitution**), and substituting the full DCT spectrum rather other only a partial frequency band of it (which we refer to as **Full Substitution**).

The I2I results of different designs of FBS are displayed in Fig. 12. It shows that Once Substitution produces severely noisy results rather than reasonable images, indicating that step-by-step FBS along the whole calibration phase is of crucial importance for smooth and stable information fusion. Removing per-step feature calibration of FBS in the early sampling process will inevitably lead to large deviation of the sampling trajectory against the reconstruction trajectory. In this case, substituting a frequency band at an intermediate time step will cause completely incoherent DCT spectrum, and thus leads to abnormal image translation results.

Besides, it also shows that Full Substitution fails to manipulate the reference image as per the text prompt. This is because substituting the full DCT spectrum is equivalent to complete feature replacement, which makes the sampling trajectory totally the same as the reconstruction trajectory during the calibration phase, the early part of the diffusion sampling process which dominates the forming of image content. Therefore, the generated image content is forced to be the same as the reference image after the calibration phase and is difficult to be modified noticeably during the subsequent non-calibration phase, the latter part of the diffusion sampling process that focuses on refining fine-grained image details rather than coarse-grained image content.

4.3 Quantitative Evaluations

For quantitative method evaluation, we evaluate methods separately on the text-driven I2I task pursuing image appearance consistency and the task pursuing image appearance divergence. For the former task, we assess models' appearance and layout preservation ability by measuring Structure Similarity (↑), Perceptual Similarity (↑), and Style Distance (↓) between I2I translation pairs. For the latter task, we assess models' contour preservation and appearance alteration capability by measuring Structure Similarity (↑) and Style Distance (↑) between I2I translation pairs. For Structure Similarity measurement, we use DINO-ViT self-similarity distance [38] as the metric for Structure Distance between two images, and define Structure Similarity as 1 - Structure Distance. We use LPIPS [42] metric to measure Perceptual Similarity, and use AdaIN style loss [14] to measure Style Distance between I2I pairs. Besides, CLIP Similarity (↑) is used to measure semantic consistency between the target text and the translated image, i.e., text fidelity of the I2I translation results. Finally, we evaluate Aesthetic Score (↑) of the translated images via the pre-trained LAION Aesthetics Predictor V2 model.

We sample reference images from LAION Aesthetics 6.5+ dataset for quantitative evaluation. For the above-mentioned two tasks, we separately sample 500 reference images for each task and manually design 2 editing text prompts for each reference image, resulting in 1000 evaluation samples (reference image and target text pairs) for each task. For evaluation of our method, we use low-FBS for the task pursuing appearance consistency and use high-FBS for the task pursuing appearance divergence. The average values of all the metrics are reported in Tab. 1. Our method achieves top rankings for all the metrics in both two tasks, indicating superiority of our method in layout and appearance preservation with low-FBS, as well as simultaneous contour preservation and appearance modification with high-FBS. Moreover, the competitive results in CLIP Similarity and Aesthetic Score reflect that our method can generate I2I results with high text fidelity and visual quality.

5 Conclusion

This paper proposes FBSDiff, a plug-and-play method adapting pre-trained T2I diffusion model to highly controllable text-driven I2I translation. At the heart of our method is decomposing different guiding factors of the reference image in the diffusion feature DCT space, and dynamically transplanting a certain DCT frequency band from diffusion features along the reconstruction trajectory into the corresponding features along the sampling trajectory, which is realized via our proposed frequency band substitution layer. Experiments demonstrate that our method allows flexible control over both guiding factors and guiding intensity of the reference image to the T2I generated image. In summary, our FBSDiff provides a novel solution to text-driven I2I translation from a frequency-domain perspective, integrating advantages in versatility, high controllability, high visual quality, and plug-and-play efficiency.

6 Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62332010, and in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

References

[1] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. 2023. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 12608–12618.

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

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Proceedings of the Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[4] Mu Cai, Hong Zhang, Huijuan Huang, Qichuan Geng, Yixuan Li, and Gao Huang. 2021. Frequency domain image translation: More photo-realistic, better identity-preserving. In *Proceedings of the IEEE International Conference on Computer Vision*. 13930–13940.

[5] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. 2022. Improving diffusion models for inverse problems using manifold constraints. *Proceedings of the Advances in Neural Information Processing Systems* 35 (2022), 25683–25696.

[6] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *Proceedings of the European Conference on Computer Vision*. Springer, 88–105.

[7] Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Proceedings of the Advances in Neural Information Processing Systems* 34 (2021), 8780–8794.

[8] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. 2023. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE International Conference on Computer Vision*. 7430–7440.

[9] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 12873–12883.

[10] Arthita Ghosh and Rama Chellappa. 2016. Deep feature extraction in the DCT domain. In *Proceedings of the International Conference on Pattern Recognition*. 3536–3541.

[11] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-prompt image editing with cross attention control. *Proceedings of the International Conference on Learning Representations* (2023).

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Proceedings of the Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

[13] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[14] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.

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

[16] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2426–2435.

[17] Gihyun Kwon and Jong Chul Ye. 2022. Diffusion-based Image Translation using disentangled style and content representation. In *Proceedings of the International Conference on Learning Representations*.

[18] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. 2023. Stylediffusion: Prompt-embedding inversion for text-based editing. *arXiv preprint arXiv:2303.15649* (2023).

[19] Zhexin Liang, Zhaochen Li, Shangchen Zhou, Chongyi Li, and Chen Change Loy. 2024. Control Color: Multimodal Diffusion-based Interactive Image Colorization. *arXiv preprint arXiv:2402.10855* (2024).

[20] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Proceedings of the Advances in Neural Information Processing Systems* 35 (2022), 5775–5787.

[21] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and L Repaint Van Gool. [n. d.]. Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11461–11471.

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

[23] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4296–4304.

[24] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the International Conference on Machine Learning*. PMLR, 16784–16804.

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

[26] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE International Conference on Computer Vision*. 4195–4205.

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *Proceedings of the International Conference on Learning Representations* (2023).

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*. PMLR, 8748–8763.

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[31] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–10.

[32] 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. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Proceedings of the Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

[33] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4713–4726.

[34] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. 2024. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4733–4743.

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. *Proceedings of the International Conference on Learning Representations* (2021).

[36] Shiqi Sun, Shancheng Fang, Qian He, and Wei Liu. 2023. Design Booster: A Text-Guided Diffusion Model for Image Translation with Spatial Layout Preservation. *arXiv preprint arXiv:2302.02284* (2023).

[37] Weimin Tan, Siyuan Chen, and Bo Yan. 2023. Diffss: Diffusion model for few-shot semantic segmentation. *arXiv preprint arXiv:2307.00773* (2023).

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

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

[40] Wenbin Xie, Dehua Song, Chang Xu, Chunjing Xu, Hui Zhang, and Yunhe Wang. 2021. Learning frequency-aware dynamic network for efficient super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*. 4308–4317.

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

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 586–595.

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