

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

Xiang Gao, Jiaying Liu*, Senior Member, IEEE

arXiv:2408.00998v2 [cs.CV] 6 Aug 2024

Abstract—Large-scale text-to-image diffusion models have been a revolutionary milestone in the evolution of generative AI and multimodal technology, 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 any model training, model 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. The code is publicly available at: <https://github.com/XiangGao1102/FBSDiff>.

Index Terms—Diffusion model, image-to-image translation, text-driven image translation.

I. INTRODUCTION

As a typical application of the booming multimodal technology, 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 [1] 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 [2] pioneers text-driven image translation by optimizing VQGAN [3] latent image embedding with CLIP text-image similarity loss. DiffusionCLIP [4] fine-tunes diffusion model [5] under the same CLIP loss to manipulate an image as

Accepted by ACMMM 2024. Xiang Gao is with Wangxuan Institute of Computer Technology, Peking University, Beijing, 100871 China, e-mail: (gaoxiang1102@pku.edu.cn). Jiaying Liu is with Wangxuan Institute of Computer Technology, Peking University, Beijing, 100871 China, e-mail: (lijiaying@pku.edu.cn). * is the corresponding author.

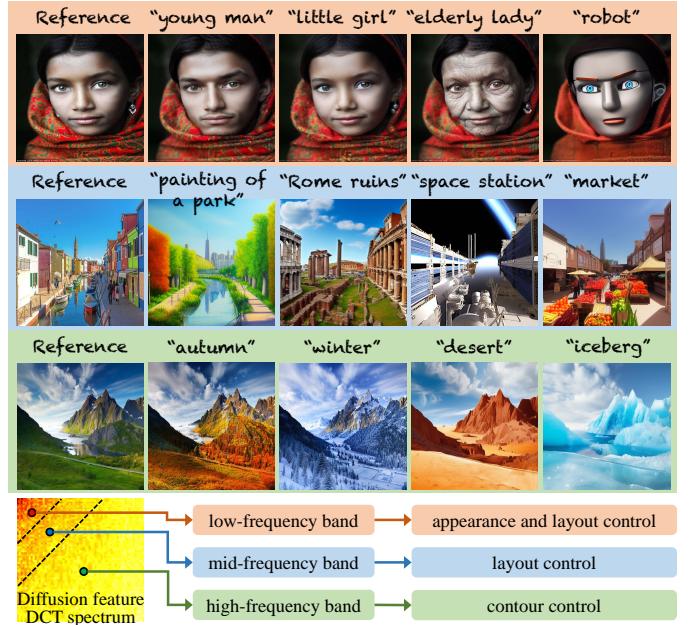


Fig. 1. Based on the pre-trained text-to-image diffusion model, FBSDiff enables efficient text-driven image-to-image translation by proposing a plug-and-play reference image guidance mechanism. It allows flexible control over different guiding factors (e.g., image appearance, image layout, image contours) 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. Better viewed with zoom-in.

per a text. DiffuseIT [6] combines VIT-based structure loss [7] and CLIP-based semantic loss to guide diffusion model's reverse sampling process via manifold constrained gradient [8], 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 inherent instability 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. Instruct-Pix2Pix [9] employs GPT-3 [10] and Stable Diffusion [11] 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 [12] trains a latent diffusion model [11] 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 [13] and Imagic [14] 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 diffusion model for each time of image manipulation, 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 [15], which severely impairs the correlation between the reference image and the translated image. To remedy this issue, Null-text Inversion [16] optimizes the unconditional null-text embedding to calibrate the sampling trajectory step by step. Prompt Tuning Inversion [17] 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 [19] 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 [20] 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) [21] proposes to leverage 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, and 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: our method extends pre-trained T2I diffusion model to the realm of I2I in a plug-and-play manner; (II) conciseness: our method 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: our method 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: our method 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. The effectiveness of our method is fully demonstrated both qualitatively and quantitatively. To summarize, we make the following key contributions:

- We provide new insights about controllable diffusion 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 modulations, 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.

II. RELATED WORK

A. Diffusion Model

Since the advent of DDPM [5], diffusion model has soon dominated the family of generative models [22]. Afterwards, much progress have been made to improve diffusion model in both methodology and application. DDIM [23] and its variants [24], [24] accelerate diffusion model sampling process to tens of times with only marginal drop in generation quality, promoting its practicability dramatically. Palette [25] extends diffusion model from unconditional image generation to the realm of conditional image synthesis, opening the door of diffusion-based image-to-image translation. With the advancement of multimodal technology, large-scale T2I diffusion models [26], [27], [28] are proposed to generate high-resolution images

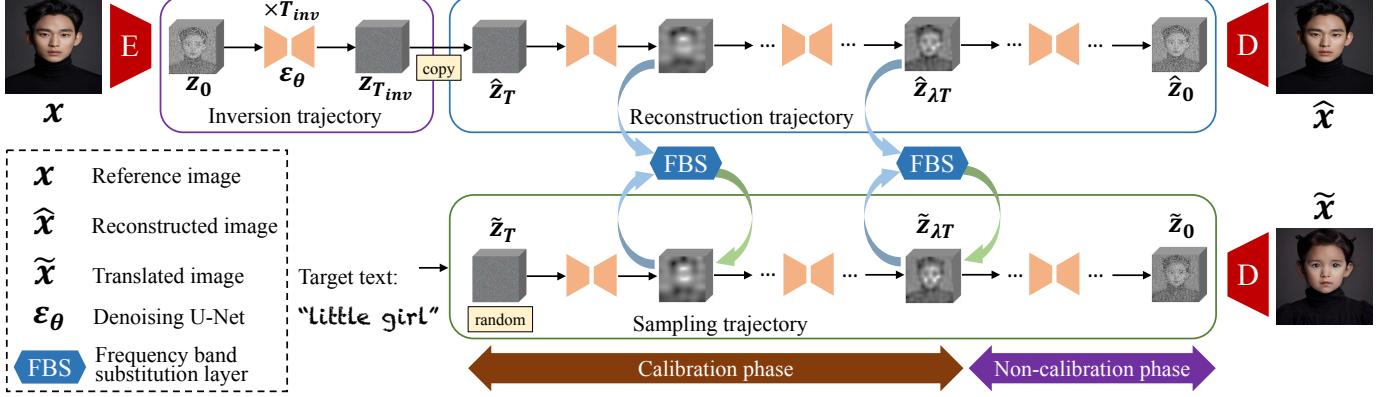


Fig. 2. Overview of FBSDiff. Based on the pre-trained latent diffusion model (LDM), FBSDiff starts with an inversion trajectory that inverts reference image into the LDM Gaussian noise space, then a reconstruction trajectory is applied to reconstruct the reference image from the inverted Gaussian noise, 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 dynamically transplanting certain DCT frequency bands from diffusion features along the reconstruction trajectory into the corresponding features along the sampling trajectory. The dynamic DCT frequency band transplantation is implemented in a plug-and-play manner with our proposed frequency band substitution layer (FBS layer).

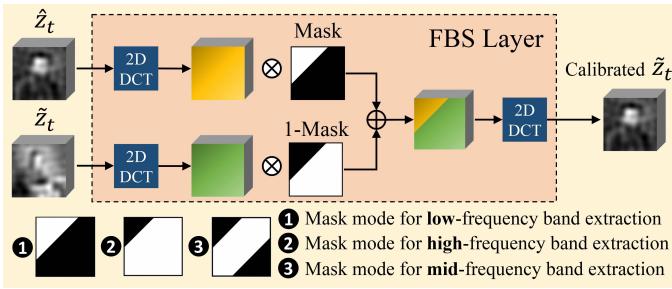


Fig. 3. Illustration of the proposed frequency band substitution (FBS) layer. The FBS layer takes in two diffusion features and substitutes a certain frequency band of one feature with the corresponding frequency band of the other feature. This is realized by converting the two diffusion features into the frequency domain via 2D DCT, extracting and transplanting a certain DCT frequency band, and converting the fused DCT features back to spatial domain via 2D IDCT. The frequency band extraction and transplantation are implemented with binary masking.

with open-domain text prompts, bringing content creation to an unprecedented level. To lower computational overhead of large-scale T2I model, Latent Diffusion Model (LDM) [11] proposes to transfer diffusion model from high-dimension pixel space to low-dimensional feature space, contributing the most widely used architecture in AIGC industry. To introduce more controllability to T2I synthesis, ControlNet [29] and T2i-adapter [30] 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 [31] and DiTs [32] propose Transformer [33] based backbone denoising network, improving T2I diffusion model to larger capacity. Up to now, diffusion model has been applied to a wide variety of vision fields such as image super-resolution [34], image inpainting [35], image colorization [36], semantic segmentation [37], point cloud generation [38], video synthesis [39], 3D reconstruction [40], etc, and is still making rapid progress in theory and potential applications.

B. Computer Vision in Frequency Perspective

Deep neural network models are mostly applied to tackle vision tasks in spatial or temporal domain, some research reveals that model performance can also be boosted from frequency domain. For example, Ghosh et al. [41] introduce DCT to convolutional neural network for image classification, accelerating network convergence speed. Xie et al. [42] propose a frequency-aware dynamic network for lightweight image super-resolution. Cai et al. [43] impose Fourier frequency spectrum consistency to image translation tasks, achieving better identity preservation ability. FreeU [44] improves T2I generation quality by selectively enhancing or depressing different frequency components of diffusion features inside the denoising U-Net model. ILVR [45] proposes to fuse low-frequency information of the forward diffusion process into the reverse sampling process for conditioned image synthesis. Our method differs with ILVR in that ILVR simulates low-pass filtering with simple feature downsampling and upsampling and performs information fusion in the spatial domain, while our method explicitly extracts and transplants frequency bands of diffusion features in the DCT domain. FCDiffusion [46] shares similar spirit of frequency-based control of T2I diffusion model with our method. However, FCDiffusion relies on training multiple frequency control branches to realize versatile control effects, while our method achieves versatility and high controllability in both guiding factor and guiding intensity of the reference image in a training-free and plug-and-play manner.

III. METHOD

In this section, we first describe the overall model architecture of our FBSDiff, then elaborate on our proposed frequency band substitution mechanism, and finally summarize our algorithm and describe implementation details. For the diffusion model background, please refer to the Appendix.

A. Overall Architecture

Established on the pre-trained Latent Diffusion Model (LDM), FBSDiff adapts it from T2I generation to the realm of text-driven I2I translation with our proposed plug-and-play reference image guidance mechanism: 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 ($\hat{z}_T \rightarrow \hat{z}_0$). 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 [5], 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 as the inversion trajectory 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 [15] by interpolating conditional (target text) and unconditional (null text) noise prediction at each time step with a guidance scale ω along the sampling process:

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

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

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

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 substitutes a certain frequency band of \tilde{z}_t in the sampling trajectory with the corresponding frequency band of \hat{z}_t in the reconstruction trajectory along the reverse sampling process. The frequency band substitution effectively and efficiently imposes guidance of the reference image to the T2I synthesis process. Both the guiding factor (e.g., image appearance, image layout, image contours) and guiding intensity of the reference image can be flexibly controlled simply by tuning the type and bandwidth of the substituted frequency band, respectively.

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., $\tilde{x} = D(\tilde{z}_0)$.

B. 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 position in the DCT spectrum of \tilde{z}_t . Finally, 2D-IDCT is applied to transform the manipulated DCT spectrum of \tilde{z}_t back into the spatial domain as the final calibrated feature.

In 2D DCT spectrum, elements with smaller coordinates (nearer to the top-left origin) encode lower-frequency information while larger-coordinate elements (nearer to the bottom-right corner) correspond to higher-frequency components. Most of the DCT spectral energy is occupied by a small proportion of low-frequency elements near the top-left origin.

In FBS layer, the sum of 2D coordinates is used as the threshold to extract DCT frequency bands of different types and bandwidths through 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 ($Mask_{mp}$):

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

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}

Algorithm 1 Complete algorithm of FBSDiff

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

- 1: Extract the initial latent feature $z_0 = E(x)$.
- 2: **for** $t = 0$ to $T_{inv} - 1$ **do**
- 3: compute z_{t+1} from z_t via Eq. 1;
- 4: **end for**{DDIM inversion}
- 5: Initialize $\hat{z}_T = z_{T_{inv}}$, $\tilde{z}_T \sim \mathcal{N}(0, I)$.
- 6: **for** $t = T$ to $\lambda T + 1$ **do**
- 7: compute \hat{z}_{t-1} from \hat{z}_t via Eq. 3;
- 8: compute \tilde{z}_{t-1} from \tilde{z}_t via Eq. 4;
- 9: substitute a certain frequency band of \tilde{z}_{t-1} with the corresponding counterpart of \hat{z}_{t-1} via Eq. 7;
- 10: **end for**{DDIM sampling in the calibration phase}
- 11: **for** $t = \lambda T$ to 1 **do**
- 12: compute \tilde{z}_{t-1} from \tilde{z}_t via Eq. 4;
- 13: **end for**{DDIM sampling in the non-calibration phase}
- 14: Obtain \tilde{z}_0 and the final translated image $\tilde{x} = D(\tilde{z}_0)$.

are respectively the lower bound and upper bound of the mid-pass filtering. Given a binary mask $Mask_* \in \{Mask_{lp}, Mask_{hp}, Mask_{mp}\}$, the frequency band substitution operation 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 the 2D-DCT and 2D-IDCT transformations respectively, which are introduced in detail in the Appendix section. The use of the 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, which controls different guiding factors of the reference image to the T2I generated result:

- **Low-frequency band** substitution enables low-frequency information guidance of the reference image x , realizing image appearance (e.g., color, luminance) and layout control over the generated image \tilde{x} ;
- **High-frequency band** substitution enables high-frequency information guidance of x , realizing image contour control over the generated image \tilde{x} ;
- **Mid-frequency band** substitution enables mid-frequency information guidance of the reference image x . By filtering out higher-frequency contour information and lower-frequency appearance information in the DCT spectrum, the substitution of the mid-frequency band realizes only image layout control over the generated image \tilde{x} .

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 x to the T2I generated image \tilde{x} .

C. 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., $\lambda=0.45$. 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.

IV. EXPERIMENTS

In this section, we first present and analyze the qualitative results of our method as well as qualitative comparison with related advanced methods; then we delve into the frequency band substitution mechanism in detail with ablation studies; finally, we show quantitative evaluations of our method and related approaches.

A. Qualitative Results

Example text-driven I2I translation results of our method are shown in Fig. 4. Our method effectively decomposes different guiding factors of the reference image by dynamically transplanting different types of DCT frequency bands of diffusion features. The low-FBS transfers low-frequency information of the reference image into the sampling trajectory, producing translated image that inherits the original image appearance and layout. In the mode of high-FBS that dynamically transplants high-frequency components of diffusion features, 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 maintain only overall image layout of the reference image, since the lower-frequency appearance information and higher-frequency contour information of the reference image are filtered out in the DCT domain. For all three modes of frequency band substitution, the image translation results exhibit high visual quality and high text fidelity for both real-world and artistic-style reference images.

The control of our method over different guiding factors of the reference image is more clearly demonstrated in Fig. 5. The T2I generated image maintains the appearance and layout of the reference image with low-FBS; preserves detailed image contours of the reference image with high-FBS; and inherits pure image layout with mid-FBS.

We qualitatively compare our method with SOTA text-driven I2I translation methods including Plug-and-Play (PAP) [21], Null-text Inversion (Null-text) [16], Pix2Pix-zero [19], InstructPix2Pix (InsPix2Pix) [9], Prompt Tuning Inversion (PT-inversion) [17], StyleDiffusion [18], and VQGAN-CLIP (VQCLIP) [2], results are displayed in Fig. 7. The top panel of Fig. 7 shows that our method with low-FBS achieves better appearance consistency between the reference image and the translated result than related approaches, and is thus better suited to image creation scenario which favors inheriting the appearance and style from an existing image.

Appearance and layout control with **low-frequency band** substitutionContour control with **high-frequency band** substitutionLayout control with **mid-frequency band** substitution

Fig. 4. Qualitative results of our method with different types of frequency band substitution. For low-frequency band substitution (low-FBS), the generated image is controlled by the reference image in terms of image appearance and layout; for high-frequency band substitution (high-FBS), the reference image controls image contours of the generated image; as for mid-frequency band substitution (mid-FBS), only image layout of the generated image is controlled by the reference image. **Better viewed with zoom-in.**

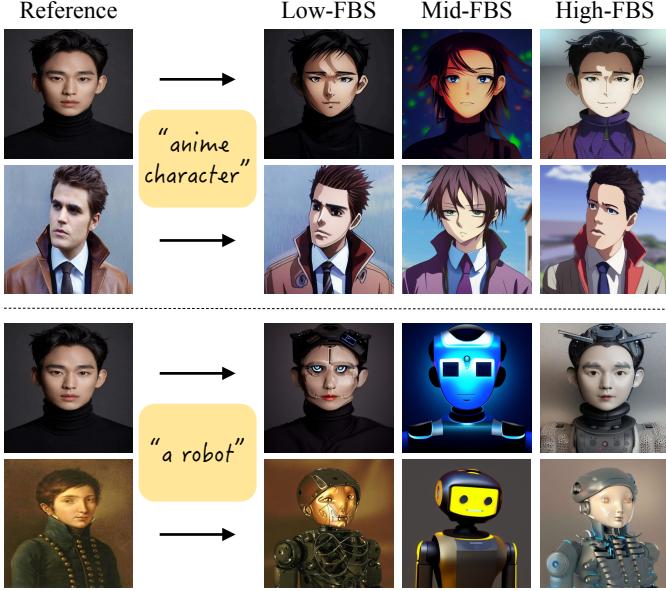


Fig. 5. Comparison among different reference image control effects achieved by low-FBS, mid-FBS, and high-FBS. Low-FBS controls image appearance and layout, mid-FBS controls only image layout, and high-FBS controls image contours.

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 images with significantly different appearance, and is thus more suitable to image creation scenario where appearance divergence is pursued. Among the compared approaches, our method is the only one that enables flexible control over different guiding factors of the reference image, and is also the only approach that simultaneously dispenses with model training, fine-tuning, online optimization, and attention modulations.

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 [16], [21], [17], [19], [18] lack such sampling diversity due to directly initializing \tilde{x}_T with the inverted feature embedding of the reference image.

The importance of frequency band substitution (FBS) for reference image control is clearly shown in Fig. 8, from which we see that low-FBS establishes appearance and layout correlations between the reference and the generated images, while removing frequency band substitution 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 the controllability in the guiding factors of the reference image, our method also allows continuous control over the guiding intensity simply by modulating the bandwidth

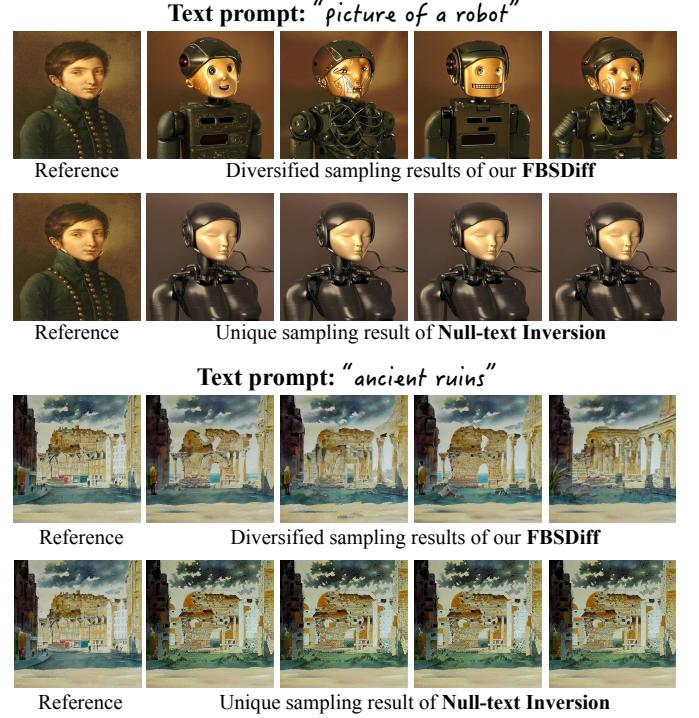


Fig. 6. Our method enables diverse sampling results for fixed reference image and text prompt, as contrasted with Null-text Inversion that produces unique text-driven I2I result. Our method also produces results with better visual quality than Null-text Inversion.

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 the image contour guiding intensity control of our method by adjusting the mid-pass filtering upper bound threshold th_{mp2} in the mode of mid-FBS. When increasing the value of th_{mp2} , more high-frequency components of the reference image (high-frequency guiding information) are included into the transplanted frequency band and transferred to the sampling trajectory, which results in more consistent image contours between the reference image and the translated image. On the contrary, decreasing the value of th_{mp2} shrinks the transplanted high-frequency guiding information and thus leads to weaker image contour consistency.

B. Ablation Study

To verify the rationality and effectiveness of our proposed method, we also explore and compare with other designs of

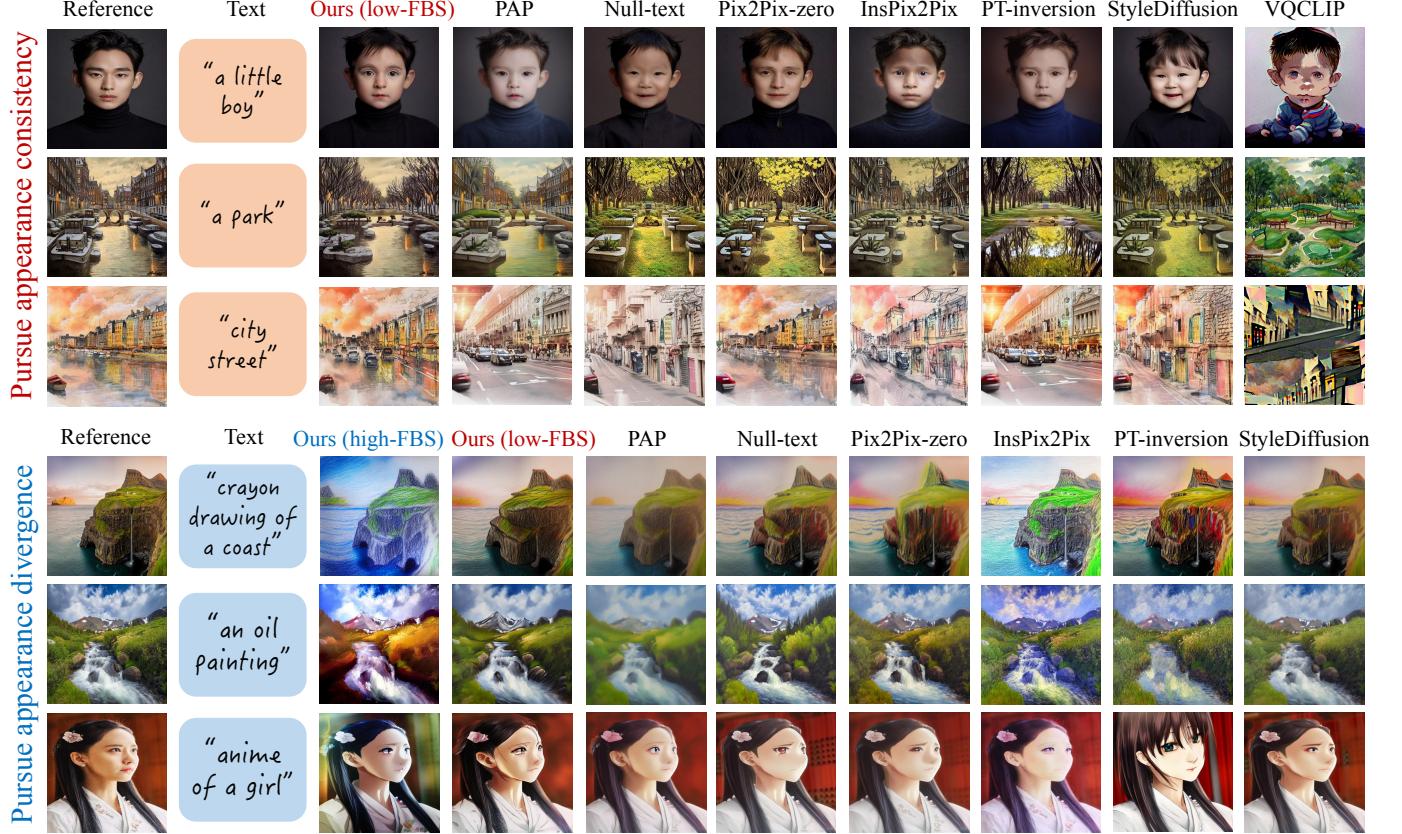


Fig. 7. Qualitative method comparisons. Our FBSDiff with low-FBS is more adept at appearance preservation than related methods, which better suits to I2I task pursuing appearance consistency between the reference image and the generated image (top panel). Conversely, our method with high-FBS remarkably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



Fig. 8. Comparison between results of our method with low-FBS and without frequency band substitution.

frequency band substitution, 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 image translation results of different designs of frequency band substitution (FBS) are displayed in Fig. 12. It shows that Once Substitution produces severely noisy results rather than reasonable images, which indicates that step-by-step FBS along the whole calibration phase is of crucial importance for smooth and stable information fusion. Since image content is basically formed in the early stage of the diffusion sampling process, 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 2D DCT spectrum, and thus leads to abnormal image translation results after converting the diffusion features back to the spatial domain.

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 that 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. Thus, the sampling

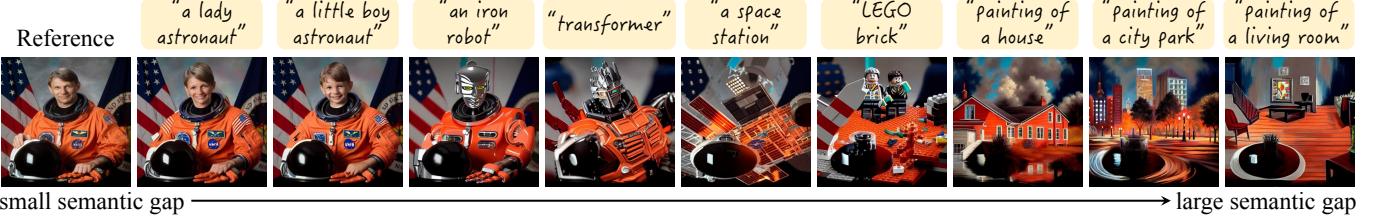


Fig. 9. Our method well adapts to varying degrees of semantic gap between the reference image and the target text prompt.

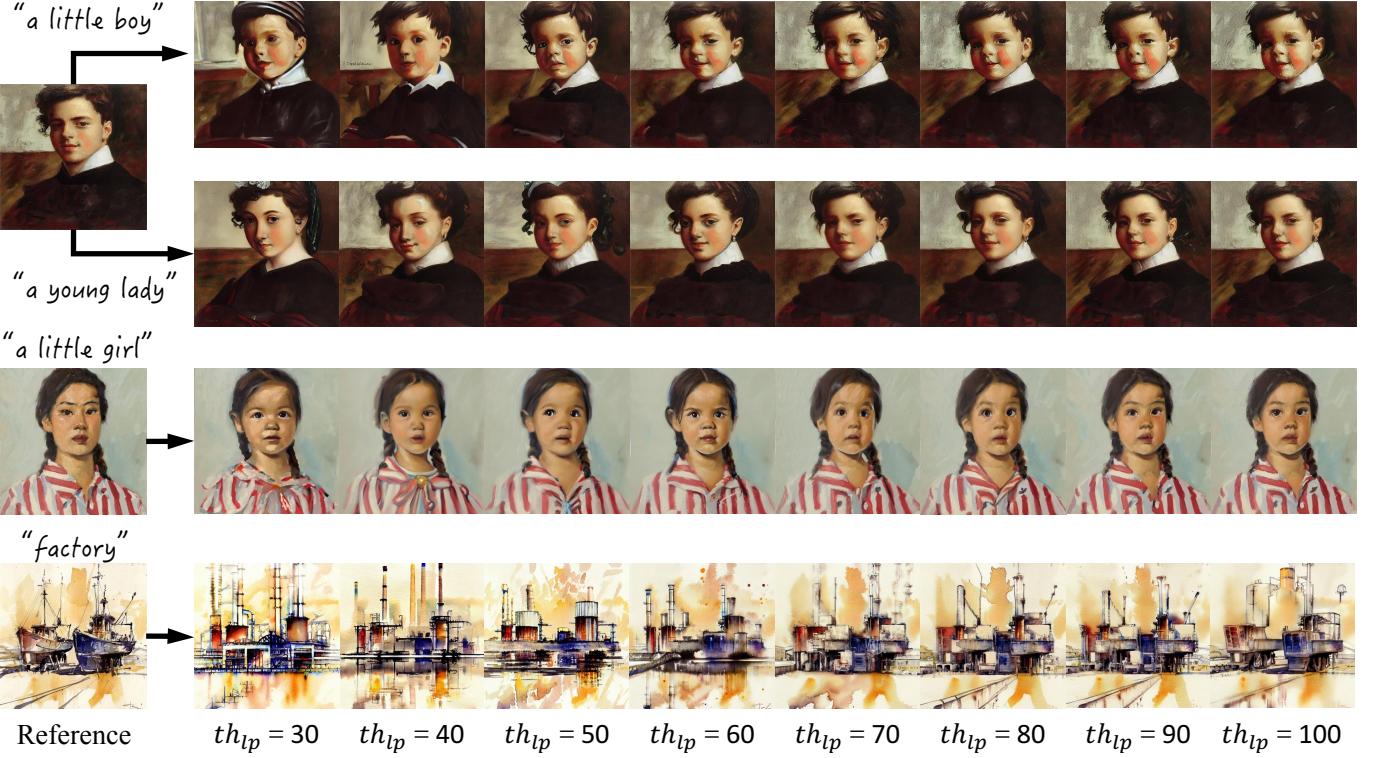


Fig. 10. Demonstration of our method in controlling the appearance and layout guiding intensity of the reference image by varying the th_{lp} in low-FBS.

results of Full Substitution closely resemble the reference images, lacking editability and text fidelity.

C. Quantitative Evaluations

For quantitative method evaluation, we separately evaluate methods on the text-driven I2I translation 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 (\uparrow), Perceptual Similarity (\uparrow), and Style Distance (\downarrow) between the reference image and the translated image pair. For the latter task, we assess models' contour preservation and appearance alteration capabilities by measuring Structure Similarity (\uparrow) and Style Distance (\uparrow) between I2I translation pairs. For Structure Similarity measurement, we use DINO-ViT self-similarity distance [7] as the metric for Structure Distance between two images, and define Structure Similarity as $1 - \text{Structure Distance}$. We use LPIPS [47] metric to measure Perceptual Similarity, and use AdaIN style loss [48] to measure Style Distance between I2I pairs. Besides, CLIP

Similarity (\uparrow) metric is used to measure semantic consistency between the target text prompt and the translated image, i.e., text fidelity of the I2I translation results. Finally, we evaluate Aesthetic Score (\uparrow) of the translated images via the pre-trained LAION Aesthetics Predictor V2 model.

We sample reference images from the 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 evaluation metrics are reported in Tab. I. 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

TABLE I
QUANTITATIVE EVALUATIONS OF TEXT-DRIVEN I2I TRANSLATION METHODS.

Emphasis		Pursuing image appearance consistency					Pursuing image appearance divergence				
Metrics	Methods	Structure Similarity(\uparrow)	LPIPS(\downarrow)	AdaIN Style Loss(\downarrow)	CLIP Similarity(\uparrow)	Aesthetic Score(\uparrow)	Structure Similarity(\uparrow)	AdaIN Style Loss(\uparrow)	CLIP Similarity(\uparrow)	Aesthetic Score(\uparrow)	
PAP [21]	0.954	0.272	20.440	0.287	6.590	0.956	28.337	0.279	6.458		
Null-text [16]	0.948	0.247	17.546	0.276	6.505	0.952	22.545	0.270	6.402		
Pix2Pix-zero [19]	0.951	0.243	16.875	0.262	6.484	0.953	21.240	0.258	6.344		
InsPix2Pix [9]	0.958	0.266	23.373	0.258	6.269	0.965	30.804	0.264	6.196		
PT-inversion [17]	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.

TABLE II
COMPARISON OF OUR APPROACH WITH RELATED MODELS IN METHOD PROPERTIES.

Methods	Training free	Fine-tuning free	Optimization free	Source-text free	Attention free	Backbone invariant
Null-text [16]	✓	✓	✗	✗	✗	✗
PAP [21]	✓	✓	✓	✗	✗	✗
Pix2Pix-zero [19]	✓	✓	✗	✗	✗	✗
InsPix2Pix [9]	✗	✓	✓	✓	✓	✓
PT-inversion [17]	✓	✓	✗	✗	✗	✗
StyleDiffusion [18]	✓	✓	✗	✗	✗	✗
VQCLIP [2]	✓	✓	✗	✓	✓	✗
DiffuseIT [6]	✓	✓	✗	✓	✓	✗
DiffusionCLIP [4]	✓	✗	✓	✓	✓	✗
Design Booster [12]	✗	✓	✓	✓	✓	✗
SINE [13]	✓	✗	✗	✗	✓	✓
Imagic [14]	✓	✗	✗	✓	✓	✓
FBSDiff (Ours)	✓	✓	✓	✓	✓	✓

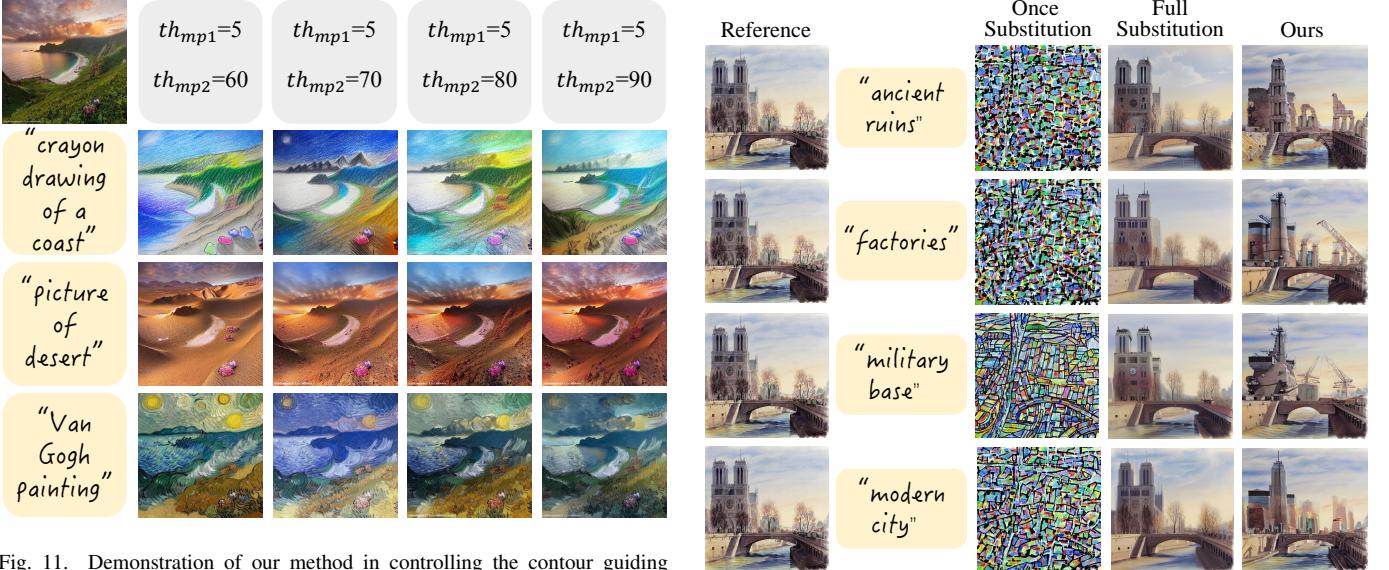


Fig. 11. Demonstration of our method in controlling the contour guiding intensity of the reference image by varying the th_{mp2} in mid-FBS.

reflect that our method can generate I2I translation results with high text fidelity and visual quality.

We compare our FBSDiff with related text-driven I2I trans-

Fig. 12. Ablation study w.r.t. different manners of frequency band substitution.

lation methods in method properties, results are summarized in Tab. II. Among the compared approaches, our method is

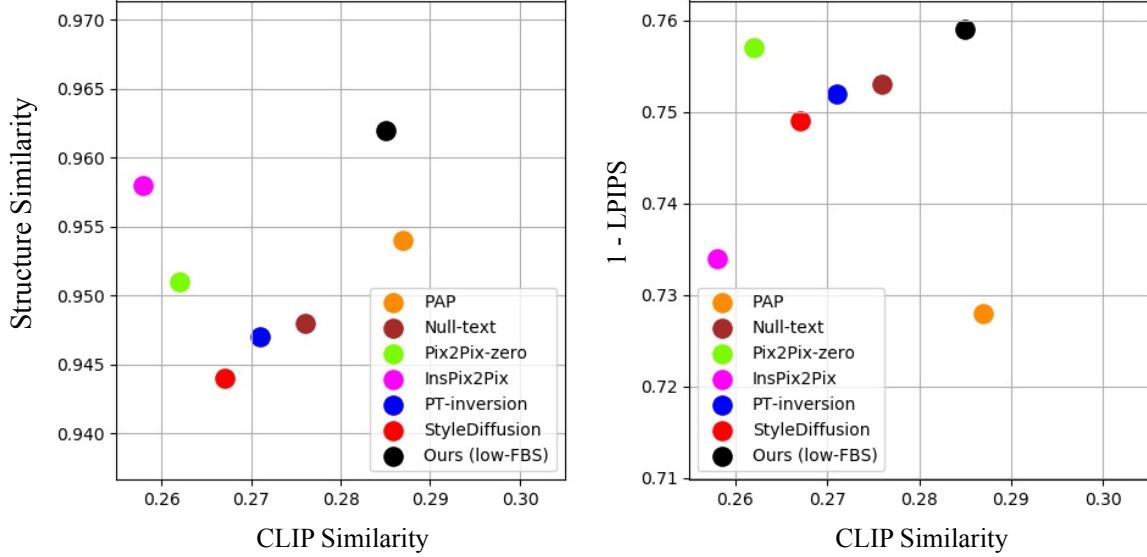


Fig. 13. Visualization of the quantitative method comparison for I2I translation task **pursuing image appearance consistency**. Left: comparison in CLIP Similarity (\uparrow) and Structure Similarity (\uparrow). Right: comparison in CLIP Similarity (\uparrow) and (1-LPIPS) (\uparrow). Our method with **low-FBS** achieves the most top-right position in both two scatters, indicating the best trade-off achieved by our method (low-FBS) in I2I translation appearance consistency and text fidelity.

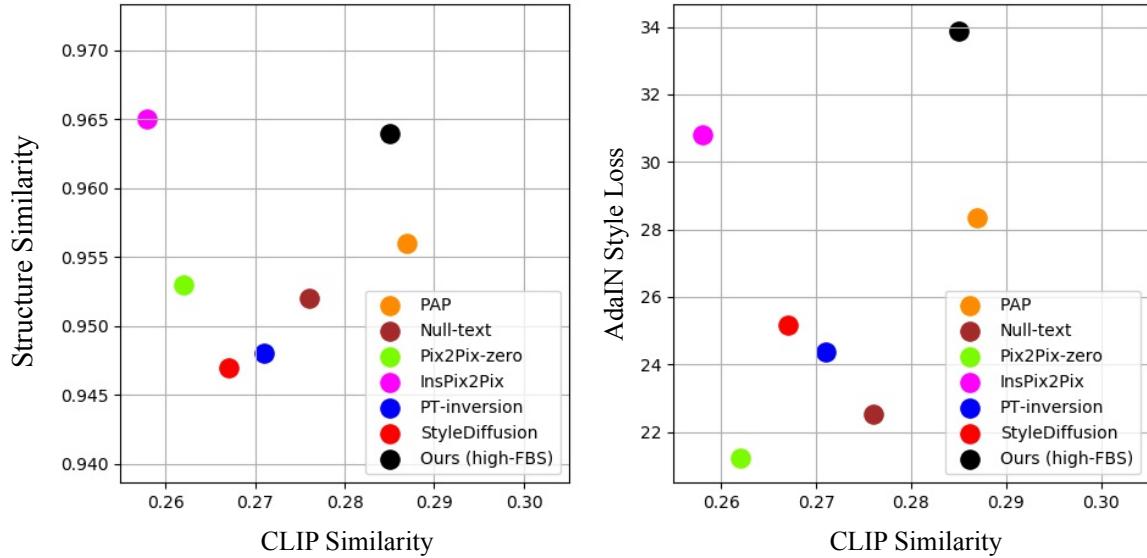


Fig. 14. Visualization of method comparison for the image translation task **pursuing image appearance divergence**. Left: comparison in CLIP Similarity (\uparrow) and Structure Similarity (\uparrow). Right: comparison in CLIP Similarity (\uparrow) and AdaIN Style Loss (\uparrow). Our method with **high-FBS** achieves the most top-right position in both two scatters, indicating the best trade-off achieved by our method (high-FBS) in I2I translation appearance divergence and text fidelity.

the only one that possesses all the following advantages:

- Dispense with model training;
- Dispense with model fine-tuning;
- Dispense with online-optimization at inference time;
- Dispense with paired source text of the reference image;
- Dispense with attention modulation operations inside the denoising network;
- Invariant to the specific architecture of the backbone diffusion model.

For quantitative evaluation reported in Tab. I, we visualize partial results to highlight the superiority of our method over

related approaches. For I2I task pursuing image appearance consistency, we display the scatter plot about Structure Similarity (\uparrow) and CLIP Similarity (\uparrow), and the scatter plot about (1-LPIPS) (\uparrow) and CLIP Similarity (\uparrow) in Fig. 13. Results show that our method with low-FBS achieves the most top-right position in both two scatter plots, indicating the best trade-off achieved by our method (low-FBS) in I2I translation appearance consistency and text fidelity. For I2I task pursuing image appearance divergence, we display the scatter plot about Structure Similarity (\uparrow) and CLIP Similarity (\uparrow), and the scatter plot about AdaIN Style Loss (\uparrow) and CLIP Similarity

(↑) in Fig. 14. Results also show the most top-right position achieved by our method with high-FBS in both two plots, indicating the best trade-off achieved by our method (high-FBS) in I2I translation appearance divergence and text fidelity.

V. 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 simply by tuning the type and bandwidth of the substituted frequency band, respectively. 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.

APPENDIX A DIFFUSION MODEL BACKGROUND

The Denoising Diffusion Probabilistic Model (DDPM) is a latent variable model that comprises a forward noising diffusion process and a reverse denoising diffusion process. Starting with a given data distribution $x_0 \sim q(x_0)$, the forward diffusion process employs a T-step Markov chain to repeatedly add Gaussian noise to the original data x_0 according to $q(x_t|x_{t-1})$ defined as follows:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1 - \alpha_t)\mathbf{I}), \quad (8)$$

where $\alpha_t \in (0, 1)$, and $\alpha_t \geq \alpha_{t+1}$. Using the notation $\bar{\alpha}_t := \prod_{i=1}^t \alpha_i$, we can derive the marginal distribution $q(x_t|x_0)$ as follows:

$$q(x_t|x_0) := \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}), \quad (9)$$

where $\sqrt{\bar{\alpha}_T}$ approaches to 0. With the above forward noising diffusion process, the source data distribution will be transformed into an isotropic Gaussian distribution.

The reverse denoising diffusion process conversely converts the isotropic Gaussian distribution to the data distribution by gradually estimating and sampling from the posterior distribution $q(x_{t-1}|x_t)$. However, $q(x_{t-1}|x_t)$ is difficult to estimate while $q(x_{t-1}|x_t, x_0)$ is tractable with some algebraic manipulation:

$$q(x_{t-1}|x_t, x_0) := \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\mathbf{I}), \quad (10)$$

$$\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}x_t, \quad (11)$$

$$\tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\beta_t, \quad (12)$$

where $\beta_t := 1 - \alpha_t$. Though no x_0 is available at inference time, its approximate value can be estimated according to Eq. 9:

$$y_\theta(x_t) := \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t)), \quad (13)$$

where $\epsilon_\theta(x_t)$ is the prediction of the Gaussian noise sampled at time step t estimated by the denoising network ϵ_θ , $y_\theta(x_t)$ is the calculated approximation of x_0 .

For image-to-image translation or text-to-image generation, additional condition (could be an image or a text) is required for noise prediction. In these cases, Eq. 13 can be updated as follows:

$$y_\theta(x_t, c) := \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t, c)), \quad (14)$$

where c denotes the additional condition that is involved in the noise prediction and the reverse denoising process.

APPENDIX B DCT AND IDCT DETAILS

We perform 2D-DCT to project diffusion features \mathbf{z} into the 2D DCT space, obtaining its frequency-domain counterpart \mathbf{f} (Eq. 15). Conversely, we employ 2D-IDCT to transform diffusion features from the DCT domain back into the spatial domain (Eq. 17). The specific form of 2D-DCT and 2D-IDCT are respectively given by Eq. 16 and Eq. 18, in which $f^{(n)}$ and $z^{(n)}$ denote the n^{th} channel of \mathbf{f} and \mathbf{z} respectively; i, j and u, v are two-dimensional coordinate indices of the spatial domain and DCT frequency domain respectively; h and w denote the height and width of the latent diffusion features; $m(0) = \frac{1}{\sqrt{2}}$, $m(\gamma) = 1$ for all $\gamma > 0$. It is worth mentioning that though the 2D-DCT and 2D-IDCT are performed on each individual channel of diffusion features (per-channel transformation), our PyTorch implementation with efficient GPU parallel computing capability enables to transform all channels simultaneously, and thus brings negligible additional time overhead during the sampling process.

$$\mathbf{f} = DCT(\mathbf{z}), \quad (15)$$

$$f_{u,v}^{(n)} = \frac{2}{\sqrt{hw}}m(u)m(v) \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} [z_{i,j}^{(n)} \cos(\frac{(2i+1)u\pi}{2h}) \cos(\frac{(2j+1)v\pi}{2w})], \quad (16)$$

$$\mathbf{z} = IDCT(\mathbf{f}), \quad (17)$$

$$z_{i,j}^{(n)} = \frac{2}{\sqrt{hw}} \sum_{u=0}^{h-1} \sum_{v=0}^{w-1} [m(u)m(v)f_{u,v}^{(n)} \cos(\frac{(2i+1)u\pi}{2h}) \cos(\frac{(2j+1)v\pi}{2w})]. \quad (18)$$

More I2I results with low-FBS (example 1)

Reference



“little boy”



“little girl”



“elderly man”



“robot”

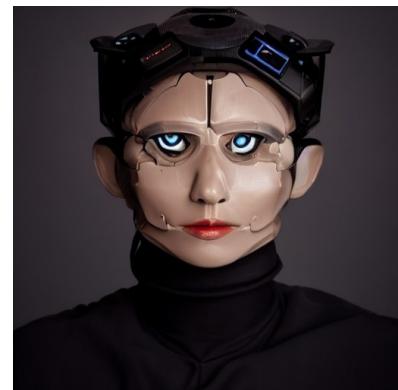


Fig. 15. More text-driven I2I results of our method with low-FBS for image appearance and layout control.

More I2I results with low-FBS (example 2)

Reference



"a boy"



"young lady"



"elderly man"



"soldier"



Fig. 16. More text-driven I2I results of our method with low-FBS for image appearance and layout control.

More I2I results with low-FBS (example 3)

Reference



"young lady"



"elderly woman"



"little boy"



"elderly man"

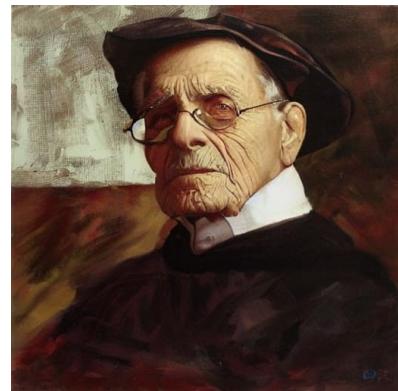
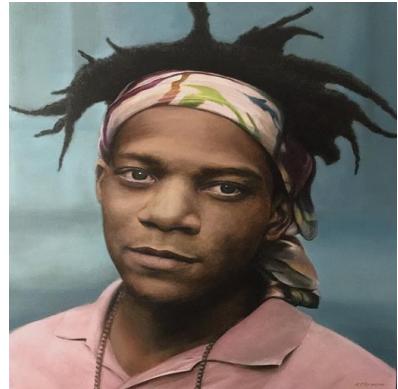


Fig. 17. More text-driven I2I results of our method with low-FBS for image appearance and layout control.

More I2I results with low-FBS (example 4)

Reference



"little boy"



"young lady"



"elderly lady"



"robot"



Fig. 18. More text-driven I2I results of our method with low-FBS for image appearance and layout control.

More I2I results with low-FBS (example 5)

Reference



"castle"



"ancient ruins"



"modern house"



"Chinese ancient buildings"



Fig. 19. More text-driven I2I results of our method with low-FBS for image appearance and layout control.

More I2I results with low-FBS (example 6)

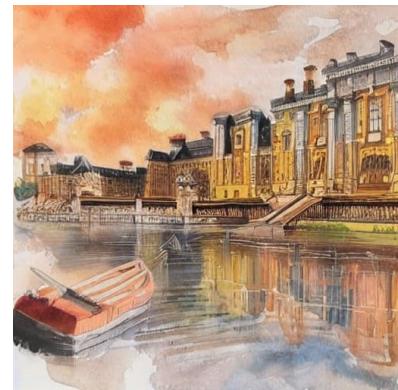
Reference



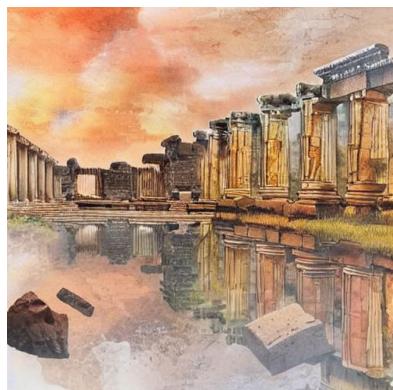
“city street”



“royal palace”



“ancient ruins”



“train”

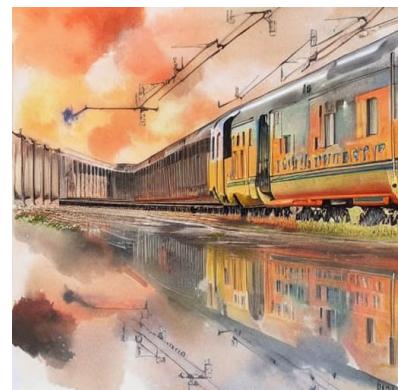


Fig. 20. More text-driven I2I results of our method with low-FBS for image appearance and layout control.

More I2I results with low-FBS (example 7)

Reference



“castle”



“city street”



“temple”



“train station”



Fig. 21. More text-driven I2I results of our method with low-FBS for image appearance and layout control.

More I2I results with low-FBS (example 8)

Reference



"castle"



"city street"



"factories"



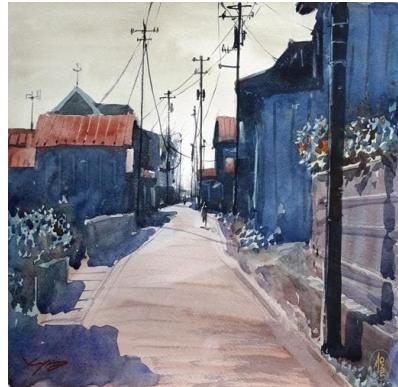
"amusement park"



Fig. 22. More text-driven I2I results of our method with low-FBS for image appearance and layout control.

More I2I results with low-FBS (example 9)

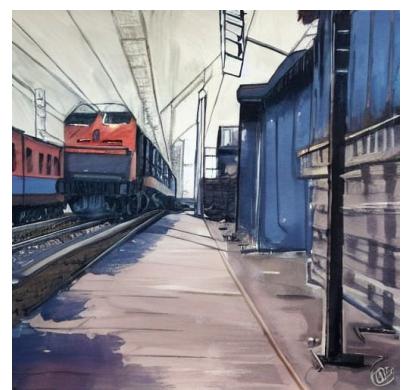
Reference



"car race"



"train"



"modern city"



"mountain landscape"



Fig. 23. More text-driven I2I results of our method with low-FBS for image appearance and layout control.

More I2I results with low-FBS (example 10)

Reference



"park"



"ruins"



"royal palace"



"ancient Rome"

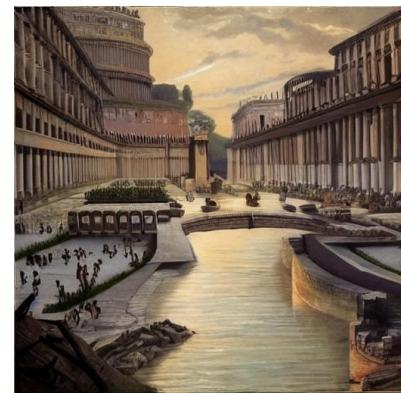


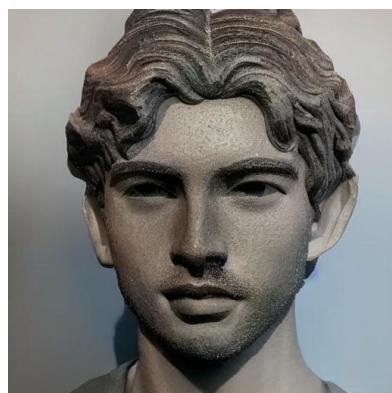
Fig. 24. More text-driven I2I results of our method with low-FBS for image appearance and layout control.

More I2I results with high-FBS (example 1)

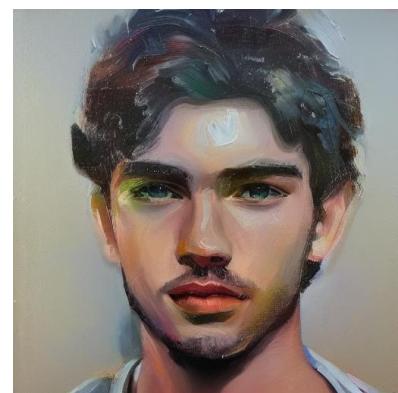
Reference



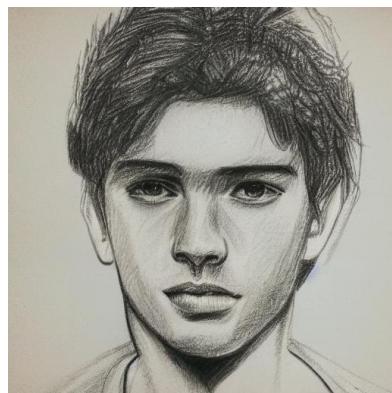
"stone sculpture"



"oil painting"



"pencil sketch"



"Van Gogh painting"

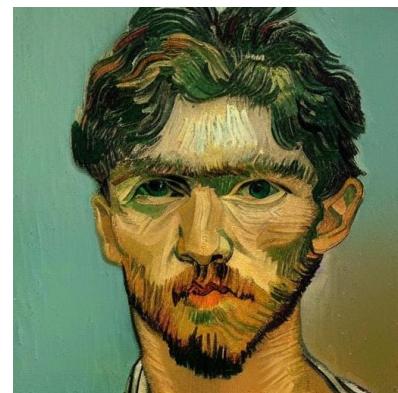


Fig. 25. More text-driven I2I results of our method with high-FBS for image contour control.

More I2I results with high-FBS (example 2)

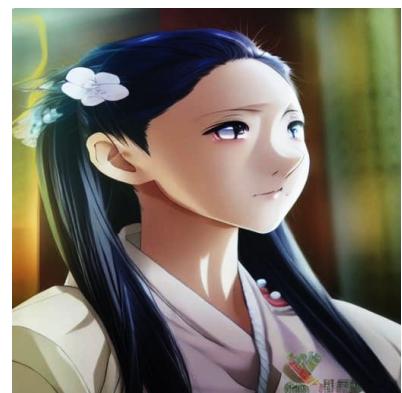
Reference



"water color"



"anime figure"



"pencil sketch"



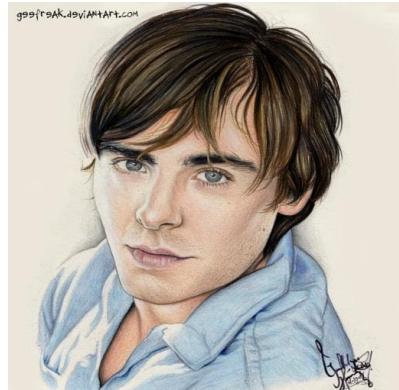
"sculpture"



Fig. 26. More text-driven I2I results of our method with high-FBS for image contour control.

More I2I results with high-FBS (example 3)

Reference



“robot”



“plaster model”



“pencil sketch”



“wooden sculpture”



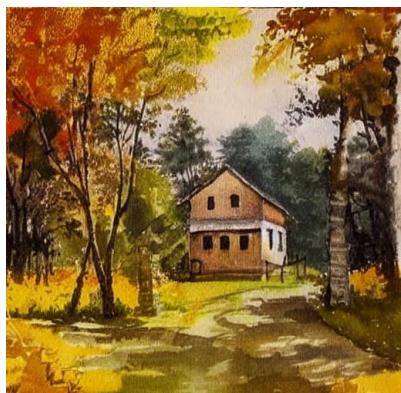
Fig. 27. More text-driven I2I results of our method with high-FBS for image contour control.

More I2I results with high-FBS (example 4)

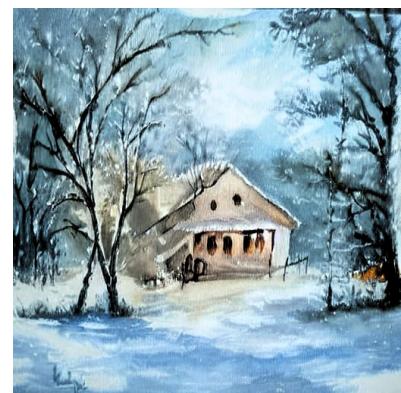
Reference



"autumn"



"winter"



"crayon drawing"



"ink-wash painting"

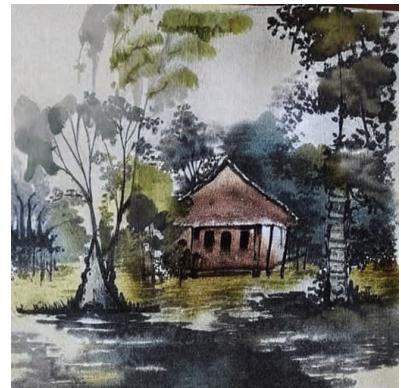


Fig. 28. More text-driven I2I results of our method with high-FBS for image contour control.

More I2I results with high-FBS (example 5)

Reference



“autumn”



“winter”



“crayon drawing”



“water color”



Fig. 29. More text-driven I2I results of our method with high-FBS for image contour control.

More I2I results with high-FBS (example 6)

Reference



“water color”



“iceberg”



“colored pencil sketch” “desert landscape”



Fig. 30. More text-driven I2I results of our method with high-FBS for image contour control.

More I2I results with high-FBS (example 7)

Reference



"autumn"



"winter"



"oil painting"



"Chinese landscape
painting"

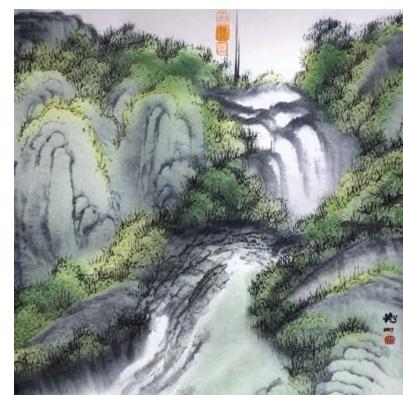
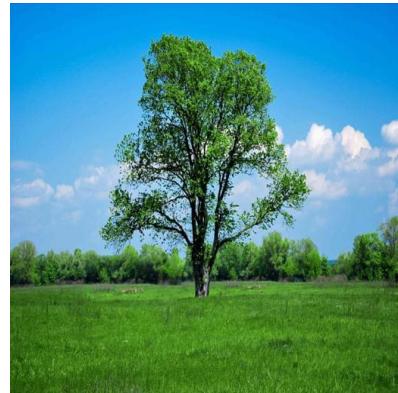


Fig. 31. More text-driven I2I results of our method with high-FBS for image contour control.

More I2I results with high-FBS (example 8)

Reference



“autumn”



“winter”



“oil painting”



“crayon drawing”



Fig. 32. More text-driven I2I results of our method with high-FBS for image contour control.

More I2I results with high-FBS (example 9)

Reference



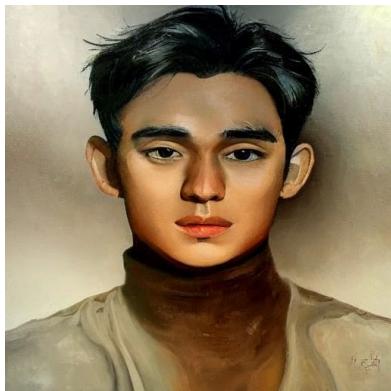
"water color"



"pencil sketch"



"oil painting"



"cartoon anime"



Fig. 33. More text-driven I2I results of our method with high-FBS for image contour control.

More I2I results with high-FBS (example 10)

Reference



"desert"



"icebergs"



"Van Gogh painting"



"Monet painting"

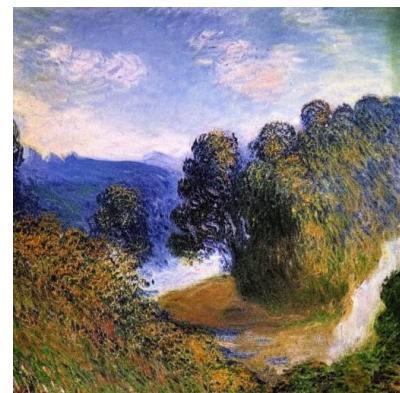


Fig. 34. More text-driven I2I results of our method with high-FBS for image contour control.

More I2I results with mid-FBS (example 1)

Reference



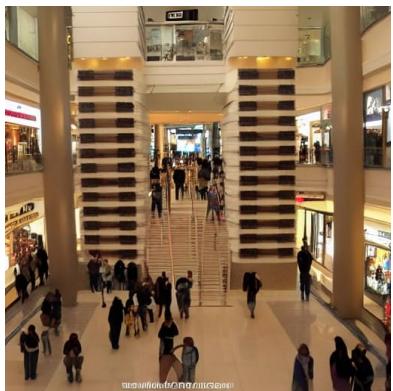
"science lab"



"bed room"



"shopping mall"



"gym"

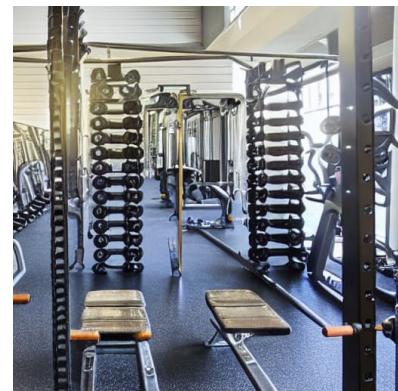


Fig. 35. More text-driven I2I results of our method with mid-FBS for image layout control.

More I2I results with mid-FBS (example 2)

Reference



"train"



"museum"



"aquarium"



"science lab"



Fig. 36. More text-driven I2I results of our method with mid-FBS for image layout control.

More I2I results with mid-FBS (example 3)

Reference



"ruins"



"lake"



"railway track"



"science lab"

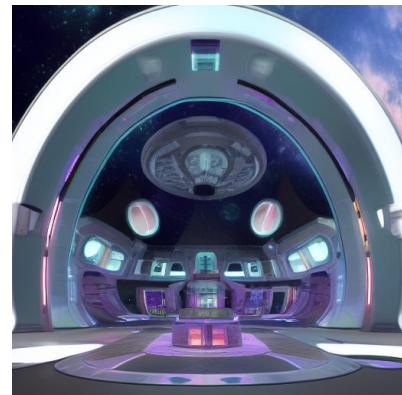
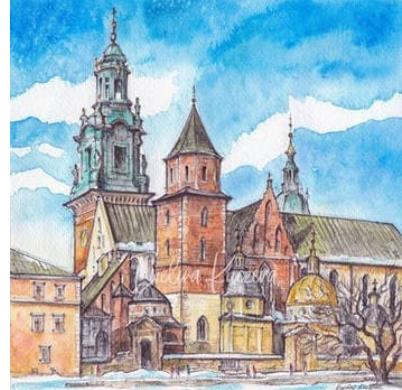


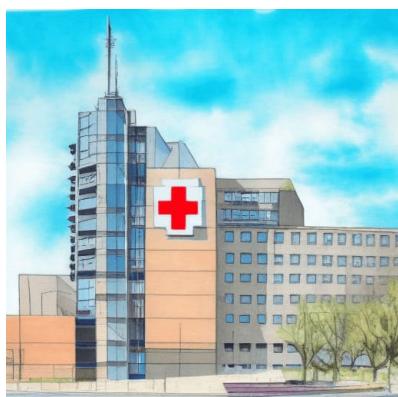
Fig. 37. More text-driven I2I results of our method with mid-FBS for image layout control.

More I2I results with mid-FBS (example 4)

Reference



“hospital”



“Big Ben”



“robots”



“rocket launching”



Fig. 38. More text-driven I2I results of our method with mid-FBS for image layout control.

More I2I results with mid-FBS (example 5)

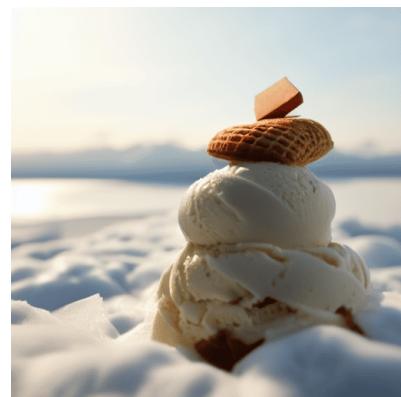
Reference



"cake"



"ice cream"



"snow man"



"LEGO brick"



Fig. 39. More text-driven I2I results of our method with mid-FBS for image layout control.

REFERENCES

- [1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *Proceedings of the International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.
- [2] K. Crowson, S. Biderman, D. Kornis, D. Stander, E. Hallahan, L. Castriato, and E. Raff, “Vqgan-clip: Open domain image generation and editing with natural language guidance,” in *Proceedings of the European Conference on Computer Vision*. Springer, 2022, pp. 88–105.
- [3] P. Esser, R. Rombach, and B. Ommer, “Taming transformers for high-resolution image synthesis,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 873–12 883.
- [4] G. Kim, T. Kwon, and J. C. Ye, “Diffusionclip: Text-guided diffusion models for robust image manipulation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2426–2435.
- [5] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Proceedings of the Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.
- [6] G. Kwon and J. C. Ye, “Diffusion-based image translation using disentangled style and content representation,” in *Proceedings of the International Conference on Learning Representations*, 2022.
- [7] N. Tumanyan, O. Bar-Tal, S. Bagon, and T. Dekel, “Splicing vit features for semantic appearance transfer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 748–10 757.
- [8] H. Chung, B. Sim, D. Ryu, and J. C. Ye, “Improving diffusion models for inverse problems using manifold constraints,” *Proceedings of the Advances in Neural Information Processing Systems*, vol. 35, pp. 25 683–25 696, 2022.
- [9] T. Brooks, A. Holynski, and A. A. Efros, “Instructpix2pix: Learning to follow image editing instructions,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 392–18 402.
- [10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, “Language models are few-shot learners,” *Proceedings of the Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.
- [11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.
- [12] S. Sun, S. Fang, Q. He, and W. Liu, “Design booster: A text-guided diffusion model for image translation with spatial layout preservation,” *arXiv preprint arXiv:2302.02284*, 2023.
- [13] Z. Zhang, L. Han, A. Ghosh, D. N. Metaxas, and J. Ren, “Sine: Single image editing with text-to-image diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6027–6037.
- [14] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, “Imagic: Text-based real image editing with diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6007–6017.
- [15] J. Ho and T. Salimans, “Classifier-free diffusion guidance,” *arXiv preprint arXiv:2207.12598*, 2022.
- [16] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. 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*, 2023, pp. 6038–6047.
- [17] W. Dong, S. Xue, X. Duan, and S. Han, “Prompt tuning inversion for text-driven image editing using diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7430–7440.
- [18] S. Li, J. van de Weijer, T. Hu, F. S. Khan, Q. Hou, Y. Wang, and J. Yang, “Styleddiffusion: Prompt-embedding inversion for text-based editing,” *arXiv preprint arXiv:2303.15649*, 2023.
- [19] G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu, “Zero-shot image-to-image translation,” in *ACM SIGGRAPH 2023 Conference Proceedings*, 2023, pp. 1–11.
- [20] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, “Prompt-to-prompt image editing with cross attention control,” *Proceedings of the International Conference on Learning Representations*, 2023.
- [21] N. Tumanyan, M. Geyer, S. Bagon, and T. 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*, 2023, pp. 1921–1930.
- [22] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *Proceedings of the Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.
- [23] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” *Proceedings of the International Conference on Learning Representations*, 2021.
- [24] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, “Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps,” *Proceedings of the Advances in Neural Information Processing Systems*, vol. 35, pp. 5775–5787, 2022.
- [25] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, “Palette: Image-to-image diffusion models,” in *ACM SIGGRAPH 2022 conference proceedings*, 2022, pp. 1–10.
- [26] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, “Glide: Towards photorealistic image generation and editing with text-guided diffusion models,” in *Proceedings of the International Conference on Machine Learning*. PMLR, 2022, pp. 16 784–16 804.
- [27] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, “Hierarchical text-conditional image generation with clip latents,” *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.
- [28] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, “Photorealistic text-to-image diffusion models with deep language understanding,” *Proceedings of the Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.
- [29] L. Zhang, A. Rao, and M. Agrawala, “Adding conditional control to text-to-image diffusion models,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.
- [30] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, and Y. Shan, “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, no. 5, 2024, pp. 4296–4304.
- [31] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, “Sdxl: Improving latent diffusion models for high-resolution image synthesis,” *Proceedings of the International Conference on Learning Representations*, 2023.
- [32] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.
- [33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Proceedings of the Advances in neural information processing systems*, vol. 30, 2017.
- [34] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, “Image super-resolution via iterative refinement,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4713–4726, 2022.
- [35] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. R. Van Gool, “Inpainting using denoising diffusion probabilistic models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11 461–11 471.
- [36] Z. Liang, Z. Li, S. Zhou, C. Li, and C. C. Loy, “Control color: Multimodal diffusion-based interactive image colorization,” *arXiv preprint arXiv:2402.10855*, 2024.
- [37] W. Tan, S. Chen, and B. Yan, “Diffss: Diffusion model for few-shot semantic segmentation,” *arXiv preprint arXiv:2307.00773*, 2023.
- [38] S. Luo and W. Hu, “Diffusion probabilistic models for 3d point cloud generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2837–2845.
- [39] S. Yu, K. Sohn, S. Kim, and J. Shin, “Video probabilistic diffusion models in projected latent space,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 456–18 466.
- [40] T. Auciukevičius, Z. Xu, M. Fisher, P. Henderson, H. Bilen, N. J. Mitra, and P. Guerrero, “Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 608–12 618.
- [41] A. Ghosh and R. Chellappa, “Deep feature extraction in the dct domain,” in *Proceedings of the International Conference on Pattern Recognition*, 2016, pp. 3536–3541.
- [42] W. Xie, D. Song, C. Xu, C. Xu, H. Zhang, and Y. Wang, “Learning frequency-aware dynamic network for efficient super-resolution,” in

- Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4308–4317.
- [43] M. Cai, H. Zhang, H. Huang, Q. Geng, Y. Li, and G. Huang, “Frequency domain image translation: More photo-realistic, better identity-preserving,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13930–13940.
 - [44] C. Si, Z. Huang, Y. Jiang, and Z. Liu, “Freeu: Free lunch in diffusion u-net,” pp. 4733–4743, 2024.
 - [45] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, “Ilvr: Conditioning method for denoising diffusion probabilistic models,” *arXiv preprint arXiv:2108.02938*, 2021.
 - [46] X. Gao, Z. Xu, J. Zhao, and J. Liu, “Frequency-controlled diffusion model for versatile text-guided image-to-image translation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 1824–1832.
 - [47] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
 - [48] X. Huang and S. Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.