Frequency-Controlled Diffusion Model for Versatile Text-Guided Image-to-Image Translation

Xiang Gao, Zhengbo Xu, Junhan Zhao, Jiaying Liu*

Wangxuan Institute of Computer Technology, Peking University, Beijing, China
{gaoxiang1102, icey.x, liujiaying}@pku.edu.cn

Abstract

Recently, text-to-image diffusion models have emerged as a powerful tool for image-to-image translation (I2I), allowing flexible image translation via user-provided text prompts. This paper proposes frequency-controlled diffusion model (FCDiffusion), an end-to-end diffusion-based framework contributing a novel solution to text-guided I2I from a frequency-domain perspective. At the heart of our framework is a feature-space frequency-domain filtering module based on Discrete Cosine Transform, which extracts image features carrying different DCT spectral bands to control the text-to-image generation process of the Latent Diffusion Model, realizing versatile I2I applications including style-guided content creation, image semantic manipulation, image scene translation, and image style translation. Different from related methods, FCDiffusion establishes a unified text-driven I2I framework suiting diverse I2I application scenarios simply by switching among different frequency control branches. The effectiveness and superiority of our method for text-guided I2I are demonstrated with extensive experiments both qualitatively and quantitatively. Our project is publicly available at: https://xianggao1102.github.io/FCDiffusion/.

Introduction

Image-to-image translation (I2I) is an attractive computer vision problem. Early I2I methods learn a cross-domain I2I mapping via GANs (Goodfellow et al. 2014). Given paired training data of two domains, Pix2Pix (Isola et al. 2017) establishes a general conditional GAN framework for supervised I2I, derived from which more complex architectures are designed for specific applications (Jiang et al. 2019; Yi et al. 2019). Since collecting paired training data is costly or even infeasible in practice, unsupervised I2I (UI2I) methods gain rapid popularity for their ability to learn I2I mapping with unpaired data. These methods employ adversarial learning to align domain distribution, and meanwhile, preserve original image content via constraints like perceptual similarity (Gao, Zhang, and Tian 2022), cycle-consistency (Zhu et al. 2017), geometry-consistency (Fu et al. 2019), and contrastive learning (Park et al. 2020), etc. Later on, numerous methods have been proposed to improve UI2I in visual quality (Wang et al. 2022), multimodality (Lee et al. 2020), multi-domain flexibility (Choi et al. 2020), efficiency (Zhang et al. 2022), and few-shot learning ability (Pizzati, Lalonde, and de Charette 2022). However, all these methods are only able to translate images between limited domains.

Subsequently, research interest has been focused on lever-

Figure 1: Our FCDiffusion adapts Stable Diffusion to versatile text-guided I2I applications, which are realized via different frequency control. Better viewed with zoom-in.

*Corresponding author. This work is supported in part by the National Natural Science Foundation of China under Grant 62332010, in part by the Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology), and in part by the China Postdoctoral Science Foundation under Grant 2023M740077.

aging CLIP (Radford et al. 2021) to guide I2I with text. VQCLIP (Crowson et al. 2022) optimizes VQGAN (Esser, Rombach, and Ommer 2021) latent vector via CLIP to manipulate an image as per a text. DiffusionCLIP (Kim, Kwon, and Ye 2022) finetunes pre-trained diffusion model with CLIP loss for text-driven I2I. DiffuseIT (Kwon and Ye 2022) uses the derivative of CLIP loss to guide diffusion model's reverse sampling process towards generating an image describing a given text. Under CLIP supervision, Text2LIVE (Bar-Tal et al. 2022) trains a network which outputs an editing mask to manipulate the source image. These CLIP-based methods allow using free-form text to instruct image translation, extending I2I from limited domains to open domains. Nonetheless, these methods are relatively slow due to online optimization. Besides, they often involve cumbersome objective functions, which makes model tuning unfriendly.

As large-scale text-to-image diffusion models revolutionize the field of generative AI, methods have been proposed to harness their immense generative power and adapt them to text-guided I2I, where the key ingredient is to establish spatial structure consistency between source and translated images. For example, SDEdit (Meng et al. 2021) preserves image overall structure by noising an image to an intermediate diffusion step and then denoising it conditioned on the text prompt. Instructpix2pix (Brooks, Holynski, and Efros 2023) directly trains a supervised text-driven I2I mapping based on large-scale paired dataset. Prompt-to-Prompt (Hertz et al. 2022) maintains image spatial structure by injecting image-text cross-attention maps of the source image along the reverse diffusion process. Similarly, Plug-and-Play (Tumanyan et al. 2023) guides the diffusion sampling process with the source image feature maps and self-attention maps to maintain image structure.

Despite stunning success of these methods, we observe that I2I has diverse application scenarios emphasizing different correlations (e.g., style, structure, layout, contour, etc.) between source and translated images, and it is difficult for a single existing method to suit all I2I scenarios well. To this end, this paper proposes a unified framework suitable for diverse I2I applications, where the intuition comes from a novel frequency-domain perspective: the source-to-target correlation of different I2I scenarios can be associated to different frequency bands of image features in Discrete Cosine Transform (DCT) spectrum. Therefore, versatile I2I could be realized by filtering image features in DCT domain, and then using the filtered image features carrying different DCT spectral bands as control signal to control the corresponding source-to-target correlation of different I2I tasks. Specifically, we summarize different I2I applications and their relations to DCT spectral bands as follows:

- **Image style translation** aims to alter image global style while keeping the original fine structures (e.g., contours, edges) maintained. These fine structures to be preserved correspond to region of high frequency in DCT spectrum, which we term *high-frequency spectral band*.
- **Style-guided content creation** aims to recreate arbitrary image content while preserving image style. The image style mainly manifests in color and luminance features that correspond to region of very low frequency in DCT spectrum, which we term *mini-frequency spectral band*.
- **Image semantic manipulation** aims to manipulate image semantic features without altering global style and spatial structure. The preserved "style + structure" corresponds to the low-frequency region in DCT spectrum that has a wider bandwidth than the mini-frequency spectral band, we term it *low-frequency spectral band*.
- **Image scene translation** aims to transform images to a greater extent where only image layout similarity is considered. To unbind restrictions on low-frequency styles and high-frequency contours, we represent pure image layout information with the middle-frequency region in DCT spectrum, i.e., *mid-frequency spectral band*.

Accordingly, we propose frequency-controlled diffusion model (FCDiffusion) for text-guided I2I. FCDiffusion basically adopts ControlNet (Zhang and Agrawala 2023) paradigm which trains a network to control the pre-trained Latent Diffusion Model (Rombach et al. 2022), where the control signal here is the frequency-domain filtered image features that carry a certain DCT spectral band. Conditioned on the control signal, the model is trained to reconstruct the filtered-out frequency spectral components of image features with the textual information from the paired text prompt. At inference time, text-driven I2I is thus allowed by feeding in arbitrary text prompt to guide the completion of the DCT spectrum. As Fig. 1 displays, our model flexibly handles diverse I2I scenarios under different modes of frequency control. The advantages of FCDiffusion are threefold: (i) it suits versatile I2I tasks simply by applying different DCT filters to construct corresponding control signals; (ii) it integrates multiple and scalable frequency control branches, enabling flexible switching among diverse I2I applications within a single model; (iii) it is concise in learning objective, low-demanding in computational resources, and competitive in I2I visual quality.

Related Work

Diffusion Models

With the advent of DDPM (Ho, Jain, and Abbeel 2020), diffusion models have received tremendous attention and have soon dominated the field of image generation (Dhariwal and Nichol 2021). Then, efforts have been made to explore their potential in various vision problems such as super-resolution (Saharia et al. 2022b), I2I (Saharia et al. 2022a), image inpainting (Lugmayr et al. 2022), etc. Boosted by vision-language multimodal technologies, large-scale text-to-image diffusion models (e.g., Imagen, DALLE2, GLIDE) impressively promote the prosperity of AIGC industry. Soon afterward, Latent Diffusion Model (LDM) (Rombach et al. 2022) enables synthesizing high-resolution images with remarkably lower computational overhead by transferring diffusion process onto low-dimensional feature space. ControlNet (Zhang and Agrawala 2023) further facilitates controllable image generation by training a network to control LDM conditioned on certain image priors like Canny edge maps. Besides, diffusion models are also increasingly applied in vision fields like point cloud generation (Luo and

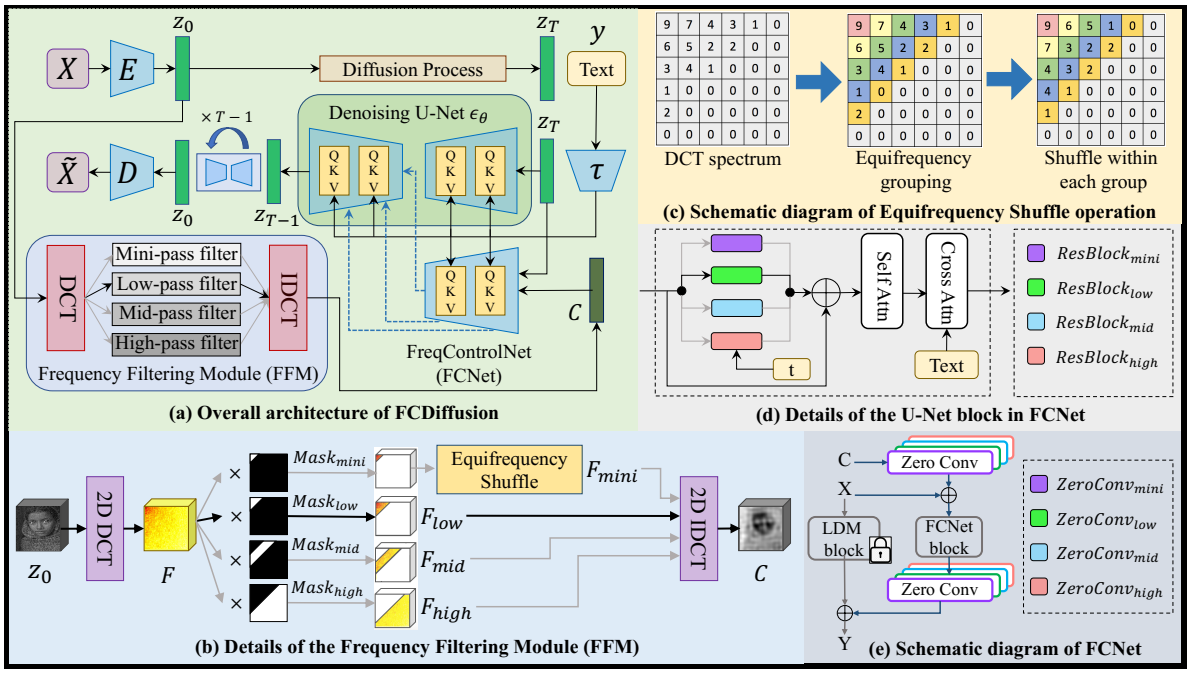

Figure 2: Overall architecture of FCDiffusion, as well as details of important modules and operations.

Hu 2021), video synthesis (Mei and Patel 2023), 3D reconstruction (Anciukevičius et al. 2023), etc.

Deep Learning in Frequency Perspective

Though neural networks are mostly exploited in spatial or temporal domains, some research work suggests that deep models can be improved from frequency-domain perspective. Ghosh et al. (Ghosh and Chellappa 2016) accelerate CNN convergence by applying DCT operation on CNN feature maps. DCT-Conv (Cheiński and Wawrzyński 2020) combines convolution with DCT inverse transformation (IDCT) to form a novel network layer taking inherent advantage in network pruning. Xie et al. (Xie et al. 2021) propose a frequency-aware dynamic network which introduces DCT to image super-resolution model to lower computation overhead. Cai et al. (Cai et al. 2021) propose to regulate image translation tasks with Fourier frequency spectrum consistency constraints, which contributes to better content preservation. This paper proposes to apply DCT filtering to text-guided I2I, specifically, to realize versatile I2I applications by extracting image features carrying different DCT frequency bands as conditional control signals.

Method

In this section, we first introduce the overall model architecture, then elaborate on important modules, and finally describe the learning objective and training details.

Overall Architecture

As illustrated in Fig. 2(a), FCDiffusion basically comprises three parts: (i) pre-trained LDM, (ii) Frequency Filtering Module (FFM), (iii) FreqControlNet (FCNet).

The pre-trained LDM uses a strong autoencoder to compress a source image $X \in R^{H\times W\times 3}$ to a compact latent representation $z_0 \in R^{h\times w\times c}$, i.e., $z_0 = E(X)$, $D(z_0) = D(E(X)) \approx X$, where $\frac{H}{h} = \frac{W}{w} = 8$, $c = 4$. A DDPM is trained at feature space to recover z_0 from Gaussian distribution conditioned on the paired text prompt y:

$$L_{LDM} = E_{z_0,y,\epsilon\sim\mathcal{N}(0,1),t}[\|\epsilon - \epsilon_\theta(z_t, t, \tau(y))\|_2^2], \quad (1)$$

where t denotes a sampled time step, z_t is the noised feature at time step t, ϵ_θ is the noise prediction U-Net which takes z_t, time step t, and text embedding $\tau(y)$ as input and outputs the estimation of the Gaussian noise sampled in the forward diffusion process, τ is the OpenCLIP transformer text encoder. We omit the technical details of DDPM and LDM here since they are not relevant to our key contributions.

To adapt LDM from text-to-image generation to text-guided I2I, a Frequency Filtering Module (FFM) is constructed to filter the encoded image features z_0 in the frequency domain, the filtered image features $C = FFM(z_0)$ function as a control signal which controls the reverse diffusion process of the LDM through a FreqControlNet (FCNet). The FCNet takes in the control signal C that contains only partial frequency spectrum components of z_0 and is optimized towards guiding the LDM to reconstruct the lossless image features z_0 with the lossy information from C and the textual information from y. From a frequency-domain perspective, the training of FCNet can be regarded as a process of recovering the filtered frequency spectrum components of z_0 via the paired text prompt y.

Frequency Filtering Module

As detailed in Fig. 2(b), in the FFM, channel-wise 2D DCT is applied to convert the initial spatial-domain image fea-

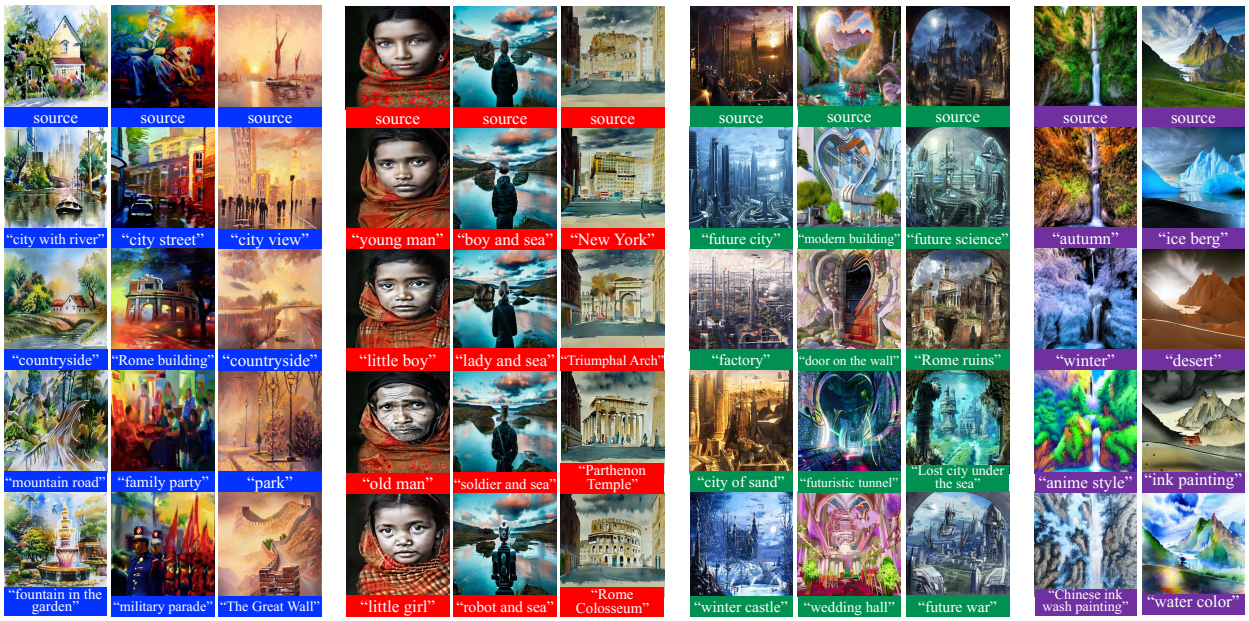

Figure 3: Example I2I results of our FCDiffusion in diverse I2I application scenarios obtained under different modes of frequency control. Better viewed with zoom-in.

tures z_0 into the frequency-domain counterpart F:

$$F_{u,v}^{(n)} = 2/(\sqrt{hw})m(u)m(v)\sum_{i=0}^{h-1}\sum_{j=0}^{w-1}[{z_0^{(n)}}_{i,j} \cos((2i+1)u\pi/(2h))\cos((2j+1)v\pi/(2w))], \quad (2)$$

where $m(0) = \frac{1}{\sqrt{2}}$, $m(\gamma) = 1$ for all $\gamma > 0$, $F^{(n)}$ and $z_0^{(n)}$ are the n^{th} channel of F and z_0 respectively, $n = 1, 2, ..., c$. u, v denote the 2D coordinates in the frequency domain. In 2D DCT spectrum, elements with smaller coordinates (nearer to the top-left origin) encode lower-frequency information, while large-coordinate elements correspond to high-frequency signals. As summarized in the Introduction part, different spectral bands in DCT domain encode different dimensions of image features and thus can be used as guiding conditions for different I2I applications. Therefore, we manually design four DCT filters (i.e., masks) for mini-pass, low-pass, mid-pass, and high-pass frequency-domain filtering respectively, as described in detail below:

$$\begin{cases} Mask_{mini}(u,v) = 1 \ \ if \ \ u+v <= 10 \ \ else \ \ 0, \\ Mask_{low}(u,v) = 1 \ \ if \ \ u+v <= 20 \ \ else \ \ 0, \\ Mask_{mid}(u,v) = 1 \ \ if \ \ 20 < u+v <= 40 \ \ else \ \ 0, \\ Mask_{high}(u,v) = 1 \ \ if \ \ u+v >= 50 \ \ else \ \ 0. \end{cases}$$

These DCT filters extract DCT features containing only the mini-frequency, low-frequency, mid-frequency, and high-frequency spectral band respectively, which are implemented via direct multiplication with F:

$$F_* = F \times Mask_*, \quad (3)$$

where $* \in \{\text{mini, low, mid, high}\}$. Finally, 2D IDCT is used to convert the filtered DCT features F_* back to the spatial domain, forming the final control signal C:

$$C_{i,j}^{(n)} = 2/(\sqrt{hw})\sum_{u=0}^{h-1}\sum_{v=0}^{w-1}[m(u)m(v){F_*^{(n)}}_{u,v} \cos((2i+1)u\pi/(2h))\cos((2j+1)v\pi/(2w))], \quad (4)$$

where $F_*^{(n)}$ and $C^{(n)}$ are the n^{th} channel of F_* and C respectively, $n = 1, 2, ..., c$. The control signal C respectively controls image style, style and structure, layout, and contour information when switching to the mini-pass, low-pass, mid-pass, and high-pass DCT filter, allowing diverse I2I scenarios emphasizing different source-to-target correlations.

Empirically, we observe that DCT features after mini-pass filtering still have some influence on the global structure of the generated image. To handle this issue and achieve pure style control, we propose and append an Equifrequency Shuffle operation to the end of the mini-pass filtering branch at inference time. As Fig. 2(c) illustrates, this operation firstly groups DCT components by the frequency level that is quantified using the sum of 2D coordinates, and then shuffles elements within each group. The Equifrequency Shuffle randomly perturbs DCT spectrum but maintains its overall energy distribution, which effectively eliminates spatial structure interference without losing style control ability.

FCNet Architectural Details

At each time step t, the FCNet takes the current denoising result z_t, together with the control signal C and the text embedding $\tau(y)$ as input, and outputs multi-scale feature maps to guide the pre-trained LDM towards reconstructing z_0. Borrowing from ControlNet (Zhang and Agrawala 2023), FCNet is a trainable copy of the LDM U-Net encoder. As Fig. 2(d) displays, each U-Net block in FCNet,

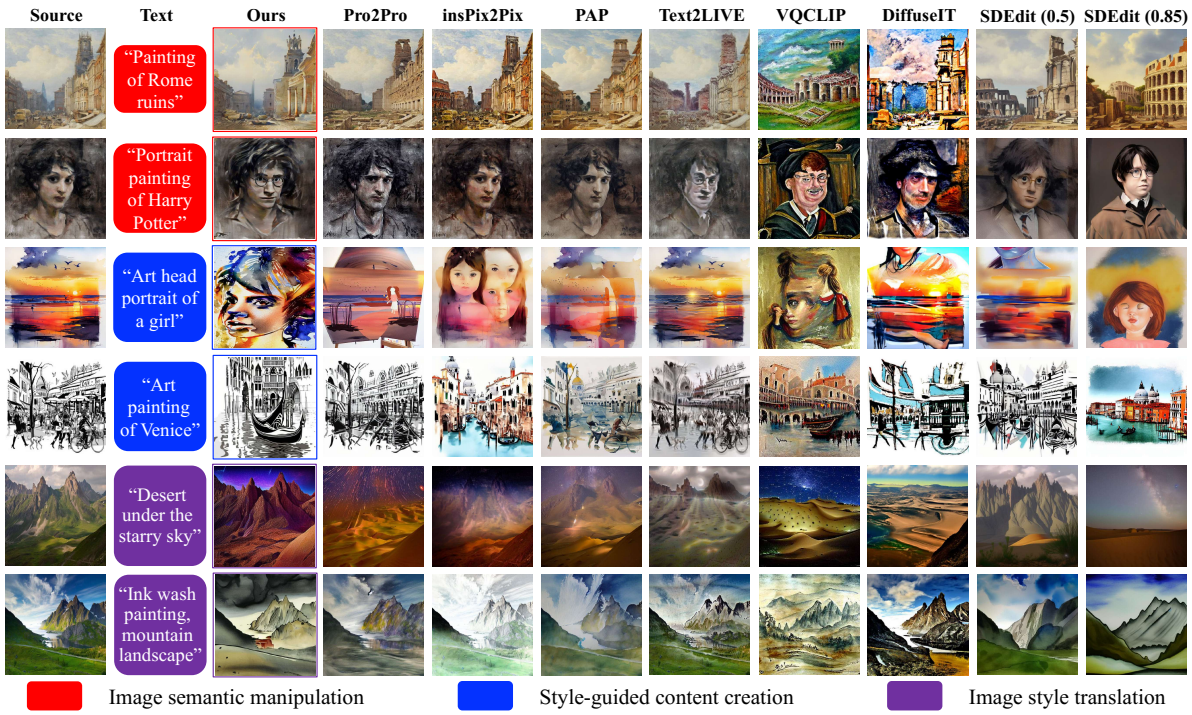

Figure 4: Visual comparisons of our method with related text-guided image translation methods on different I2I tasks including image semantic manipulation (top two rows), style-guided content creation (middle two rows), and image style translation (bottom two rows). Results of our method for these three tasks are obtained by switching to the low-frequency, mini-frequency, and high-frequency control branch respectively. Better viewed with zoom-in.

in turn, consists of a ResBlock fusing the time embedding, a self-attention block, and a cross-attention block fusing the text embedding. Each ResBlock in FCNet has four parallel replicates corresponding to the four DCT filtering branches in FFM. As Fig.2(e) illustrates, the Zero Convolutions proposed in ControlNet (Zhang and Agrawala 2023) are also utilized here for smooth feature fusion or feature injection. Similarly, each Zero Convolution also has four parallel replicates corresponding to the four DCT filtering branches.

Learning Objective and Training Details

Our framework is fully differentiable and end-to-end trainable. The objective is to reconstruct the lossless image features z_0 with the lossy control signal $C = FFM(z_0)$ and the paired text prompt y, which is equivalent to minimizing the following conditional noise regression loss:

$$L = E_{z_0,y,\epsilon\sim\mathcal{N}(0,1),t}[||\epsilon - \epsilon_\theta(z_t, t, \tau(y), c_t)||_2^2], \quad (5)$$

$$c_t = FCNet(FFM(z_0), t, \tau(y)). \quad (6)$$

The FCNet, as the only trainable module in the entire framework, is initialized from the pre-trained LDM except for the additional Zero Convolutions. In the training phase, we freeze all self-attention and cross-attention layers in the FCNet and only finetune ResBlocks and Zero Convolutions. We observe that freezing and sharing attention layers reduce trainable parameters without degrading generation performance. Our model contains four frequency control branches consisting of [Mask$_*$, ResGroup$_*$, ZeroGroup$_*$], where $* \in$ {mini, low, mid, high}. ResGroup$_*$ represents the parameter group comprising all ResBlock$_*$ replicates in the FCNet that correspond to the DCT filter Mask$_*$, and the same for ZeroGroup$_*$. These four frequency control branches are separately finetuned for different I2I scenarios and can be flexibly switched at inference time. It is worth mentioning that the control branches in our model are scalable and pluggable, more I2I control effects can be implemented simply by designing corresponding DCT filters and allocating additional ResGroup and ZeroGroup for finetuning.

Experiments

Experiment Setup

We use Stable Diffusion v2-1-base as the pre-trained LDM in our model, and use LAION-Aesthetics 6.5+ which contains 625K image-text pairs as our dataset, in which we randomly partition into a training set and a test set at a ratio of 9:1. We train at 512×512 image resolution, i.e., $H = W = 512, h = w = 64$. We set the initial learning rate as 1e-5. Each frequency control branch in our model is separately finetuned for 100K iterations with batch size 4 on a single RTX 3090 Ti GPU. At inference time, FCDiffusion can flexibly adapt to versatile I2I scenarios simply by switching among different frequency control branches. All the results in this paper are generated using the DDIM (Song, Meng, and Ermon 2020) sampler with 50 steps.

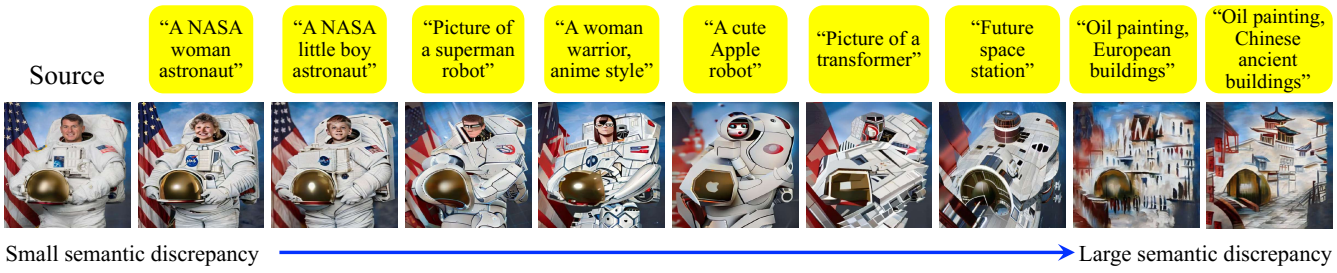

Figure 5: With low-frequency control, our method is able to manipulate image semantics under different degrees of semantic discrepancy. As the semantic gap between the source image and the target text increases, the translated image can still conform to the text with the original image style and structure preserved. Better viewed with zoom-in.

Figure 6: Qualitative ablation study on the architectural design of the FCNet. Better viewed with zoom-in.

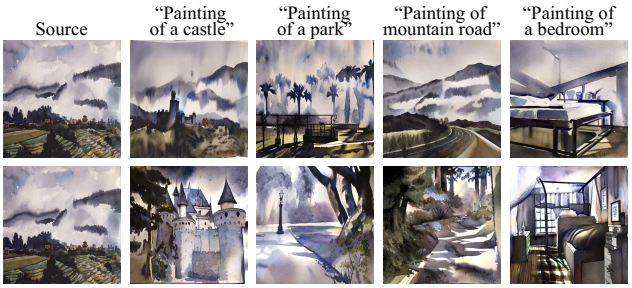

Figure 7: Qualitative ablation study on the style-guided content creation task without (top row) and with (bottom row) Equifrequency Shuffle. Better viewed with zoom-in.

Qualitative Analyses

As Fig. 3 displays, our method produces high-quality results in diverse I2I scenarios by switching among different frequency control branches. With mini-frequency control, the translated image only preserves image style with little constraint on spatial structure and thus allows style-guided content creation. Under low-frequency control, both image style and spatial structure are maintained, which suits small-scale image editing, i.e., image semantic manipulation. Under high-frequency control, the translated image conforms to the source image in object contours with little constraint on global style, which allows image style translation. Besides, we also realize image scene translation where the source-to-target correlation is the image layout, for which we resort to mid-frequency control to bypass constraints on low-frequency style and high-frequency contours.

In Fig. 4, we qualitatively compare our method with related advanced models on some challenging I2I examples. We abbreviate methods of Prompt-to-Prompt, InstructPix2Pix, Plug-and-Play as Pro2Pro, insPix2Pix, PAP respectively for brevity. For image semantic manipulation, Pro2Pro, insPix2Pix, and PAP are less able to generate semantically faithful results while achieving high consistency in global style and spatial structure. Results of Text2LIVE contain severe artifacts. VQCLIP and DiffuseIT fail to preserve original style distribution. In the track of style-guided content creation, Pro2Pro struggles to generate content faithful to the text. Results of insPix2Pix better comply with the text but are weaker in style preservation. Other methods either fail to generate faithful content or fail to maintain original style. For image style translation, results of VQCLIP manifest accurate style distribution instructed by the text but lose original contour information. Text2LIVE, on the contrary, preserves contours well but fails to precisely translate image style. Other methods either fail to maintain contour consistency or cannot translate image style faithfully.

Besides, we also compare to SDEdit with different noising strength values (shown in the parentheses). SDEdit(0.5) can maintain basic visual features of the source image, but is less effective in maintaining style distribution for image semantic manipulation, creating structure-invariant content for style-guided content creation, and altering image style sufficiently for image style translation. For large noising strength, results of SDEdit(0.85) suffer from weak connection to the source image, losing the corresponding source-to-target correlation in all tracks. By contrast, our method achieves the desired performance in all I2I scenarios.

As demonstrated in Fig. 5, our method can handle image semantic manipulation with not only narrow semantic gaps but also large semantic discrepancies based on the low-frequency control. The semantically translated images can still comply with the text and preserve the original "style+structure" simultaneously even if the target text and the source image are semantically uncorrelated.

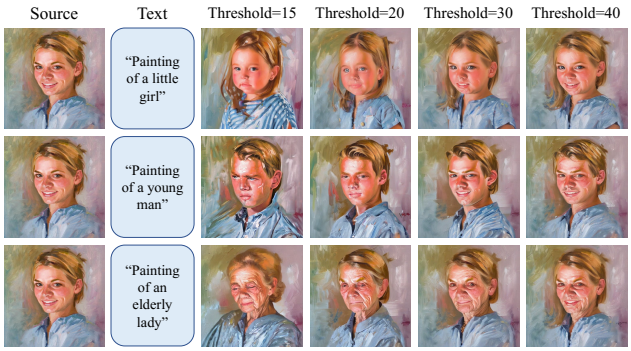

Figure 8: Example image semantic manipulation results of our method using the low-frequency control with different low-pass filtering thresholds.

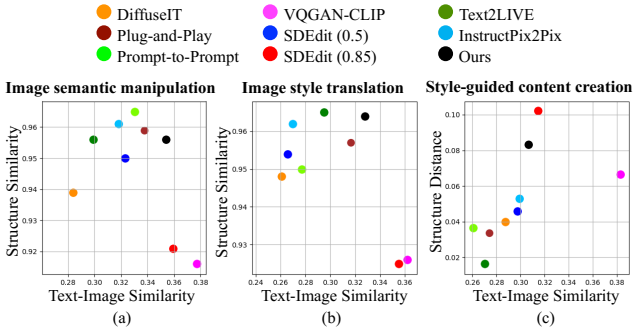

Figure 9: Quantitative comparison of different I2I methods in text fidelity and image structure preservation. We evaluate on three different I2I tracks, our method consistently achieves good trade-off between these two aspects.

Ablation Studies

We ablate our model design from the following four aspects: (i) the necessity of injecting time embedding into FCNet; (ii) the necessity of injecting text embedding into FCNet; (iii) the effectiveness of our Equifrequency Shuffle operation for image spatial structure de-correlation; (iv) the influence of the frequency band range to I2I. We conduct the first study by removing the time embedding fusion layers from all ResBlocks in FCNet. For the second study, we do not modify the cross-attention layer but instead feed the null text to the FCNet during training. As qualitatively displayed in Fig. 6, removing time embedding injection yields unstable I2I results with noticeable noises and artifacts, which means that it is important for FCNet to learn to provide time-dependent guidance information. Removing text embedding injection also leads to lower-quality results, indicating that the textual information contributes to FCNet in providing finer control to LDM. Example results of the third study are shown in Fig. 7, from which we see that the style-guided content creation results obtained without Equifrequency Shuffle still resemble the source image in global structure, whereas results obtained with Equifrequency Shuffle are much more disentangled in spatial structure, demonstrating the effectiveness of this operation in weakening structure correlation for style-guided content creation task. For the last study, we qualitatively compare the image semantic manipulation results of our method achieved under the low-frequency control with varying low-pass filtering thresholds (i.e., varying bandwidth of the extracted low-frequency spectral band). Results displayed in Fig. 8 demonstrate that the higher the threshold is, the closer the translated image is to the source image. Intuitively, raising the low-pass filtering threshold results in DCT features with wider spectral band, namely more source image information in the control signal. Conversely, narrower bandwidth corresponds to less source information and thus more variation in the generated images.

Quantitative Evaluations

For quantitative evaluation, we conduct 200 text-guided image translations for each I2I application scenario and compute the average value of the text-image similarity score and the image structure distance. The text-image similarity score is used to measure the fidelity of the translated image to the target text prompt, for which we use CLIP cosine similarity (Radford et al. 2021) as the metric. The image structure distance aims to measure the spatial structure discrepancy between the source and the translated image pairs, for which we use DINO-ViT self-similarity distance (Tumanyan et al. 2022) as the metric. Correspondingly, we define image structure similarity as 1 minus image structure distance.

As Fig. 9 shows, for image semantic manipulation and image style translation, methods are expected to achieve both high text-image similarity score and faithful preservation of source image spatial structure (i.e., large image structure similarity). Though these two metrics contradict each other, our method achieves the most top-right position in Fig. 9(a) and 9(b), indicating a good trade-off between text fidelity and structure consistency. For style-guided content creation, methods are expected to generate structure-disentangled I2I results with good compliance with the text, therefore, large text-image similarity score and large image structure distance are preferred. As Fig. 9 (c) shows, our method performs well on both two sides. Though SDEdit(0.85) and VQGAN-CLIP achieve more top-right positions in this track, these two methods suffer from weak I2I structure consistency in the former two tasks.

Conclusion

In this paper, we propose a solution to text-guided I2I from a novel frequency-domain perspective. The proposed model, FCDiffusion, adapts the pre-trained LDM from text-to-image generation to text-guided I2I by filtering the initial image features in the DCT domain and using the filtered image features carrying different DCT spectral bands to control the reverse diffusion process of LDM. Our method can be interpreted as learning to complete the full DCT spectrum with the textual information in the paired text prompt, conditioned on the filtered partial DCT spectral band. By designing different modes of DCT filters, our method realizes different control effects for text-guided I2I and enables versatile I2I application scenarios with a single model. The versatility and high-quality I2I performance of our method are comprehensively proved with extensive experiments.

References

Anciukevičius, T.; Xu, Z.; Fisher, M.; Henderson, P.; Bilen, H.; Mitra, N. J.; and Guerrero, P. 2023. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12608–12618.

Bar-Tal, O.; Ofri-Amar, D.; Fridman, R.; Kasten, Y.; and Dekel, T. 2022. Text2live: Text-driven layered image and video editing. In *Proceedings of the European Conference on Computer Vision*, 707–723.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 18392–18402.

Cai, M.; Zhang, H.; Huang, H.; Geng, Q.; Li, Y.; and Huang, G. 2021. Frequency domain image translation: More photo-realistic, better identity-preserving. In *Proceedings of the IEEE International Conference on Computer Vision*, 13930–13940.

Cheiński, K.; and Wawrzyński, P. 2020. DCT-Conv: Coding filters in convolutional networks with Discrete Cosine Transform. In *International Joint Conference on Neural Networks*, 1–6.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8188–8197.

Crowson, K.; Biderman, S.; Kornis, D.; Stander, D.; Hallahan, E.; Castricato, L.; and Raff, E. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *Proceedings of the European Conference on Computer Vision*, 88–105.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12873–12883.

Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Zhang, K.; and Tao, D. 2019. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2427–2436.

Gao, X.; Zhang, Y.; and Tian, Y. 2022. Learning to Incorporate Texture Saliency Adaptive Attention to Image Cartoonization. In *International Conference on Machine Learning*, 7183–7207.

Ghosh, A.; and Chellappa, R. 2016. Deep feature extraction in the DCT domain. In *International Conference on Pattern Recognition*, 3536–3541.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1125–1134.

Jiang, Y.; Lian, Z.; Tang, Y.; and Xiao, J. 2019. Scfont: Structure-guided Chinese font generation via deep stacked networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4015–4022.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2426–2435.

Kwon, G.; and Ye, J. C. 2022. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*.

Lee, H.-Y.; Tseng, H.-Y.; Mao, Q.; Huang, J.-B.; Lu, Y.-D.; Singh, M.; and Yang, M.-H. 2020. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128: 2402–2417.

Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; and Van Gool, L. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 11461–11471.

Luo, S.; and Hu, W. 2021. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2837–2845.

Mei, K.; and Patel, V. 2023. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 9117–9125.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.-Y.; and Ermon, S. 2021. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive learning for unpaired image-to-image translation. In *Proceedings of the European Conference on Computer Vision*, 319–345.

Pizzati, F.; Lalonde, J.-F.; and de Charette, R. 2022. Manifest: Manifold deformation for few-shot image translation. In *Proceedings of the European Conference on Computer Vision*, 440–456.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; and Norouzi, M. 2022a. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH Conference Proceedings*, 1–10.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022b. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45: 4713–4726.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tumanyan, N.; Bar-Tal, O.; Bagon, S.; and Dekel, T. 2022. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10748–10757.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 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.

Wang, T.; Zhang, T.; Zhang, B.; Ouyang, H.; Chen, D.; Chen, Q.; and Wen, F. 2022. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*.

Xie, W.; Song, D.; Xu, C.; Xu, C.; Zhang, H.; and Wang, Y. 2021. Learning frequency-aware dynamic network for efficient super-resolution. In *Proceedings of the IEEE International Conference on Computer Vision*, 4308–4317.

Yi, R.; Liu, Y.-J.; Lai, Y.-K.; and Rosin, P. L. 2019. Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10743–10752.

Zhang, L.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*.

Zhang, L.; Chen, X.; Tu, X.; Wan, P.; Xu, N.; and Ma, K. 2022. Wavelet knowledge distillation: Towards efficient image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12464–12474.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2223–2232.