

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

Anonymous Authors

1 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:

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

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:

$$q(x_t|x_0) := \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}), \quad (2)$$

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:

$$q(x_{t-1}|x_t, x_0) := \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\mathbf{I}), \quad (3)$$

$$\tilde{\mu}_t(x_t, x_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}x_t, \quad (4)$$

$$\tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\beta_t, \quad (5)$$

where $\beta_t := 1 - \alpha_t$. Though no x_0 is available at inference time, its approximate value can be estimated according to Eq. 2:

$$y_\theta(x_t) := \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t)), \quad (6)$$

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. 6 can be updated as follows:

$$y_\theta(x_t, c) := \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t, c)), \quad (7)$$

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

2 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. 8). Conversely, we employ 2D-IDCT to transform diffusion features in the DCT domain back into the spatial domain (Eq. 10). The specific form of 2D-DCT and 2D-IDCT are respectively given by Eq. 9 and

Eq. 11, in which $f^{(n)}$ and $z^{(n)}$ denote the n^{th} channel of \mathbf{f} and \mathbf{z} respectively, $n = 1, 2, 3, 4$ for our used Latent Diffusion Model backbone [10]; 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, $h = w = 64$ for generating 512×512 resolution images; $m(0) = \frac{1}{\sqrt{2}}$, $m(\gamma) = 1$ for all $\gamma > 0$. It is worth mentioning that though the 2D-DCT and 2D-IDCT are performed on each channel of diffusion features (per-channel transformation), our PyTorch implementation with efficient GPU parallel computing enables to transform all feature channels simultaneously, and thus brings negligible time overhead during the sampling process.

$$\mathbf{f} = 2D - DCT(\mathbf{z}), \quad (8)$$

$$f_{u,v}^{(n)} = \frac{2}{\sqrt{hw}}m(u)m(v) \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} [z_{i,j}^{(n)} \cos(\frac{(2i+1)u\pi}{2h}) \cos(\frac{(2j+1)v\pi}{2w})], \quad (9)$$

$$\mathbf{z} = 2D - IDCT(\mathbf{f}), \quad (10)$$

$$z_{i,j}^{(n)} = \frac{2}{\sqrt{hw}} \sum_{u=0}^{h-1} \sum_{v=0}^{w-1} [m(u)m(v)f_{u,v}^{(n)} \cos(\frac{(2i+1)u\pi}{2h}) \cos(\frac{(2j+1)v\pi}{2w})]. \quad (11)$$

3 METHOD PROPERTY SUMMARY

We compare our FBSDiff with related text-driven I2I translation methods in method properties, results are summarized in Tab. 1. Among the compared approaches, our method is the only one that possesses all the following advantages:

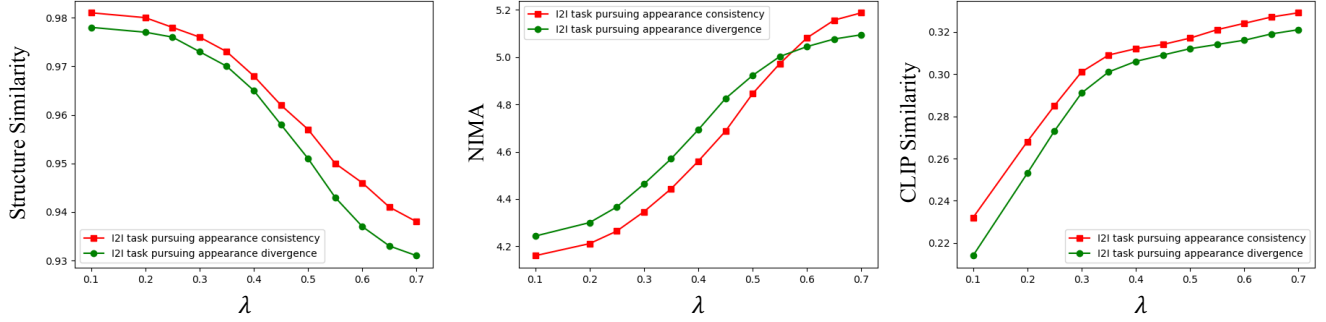
- (1) Do not require model training;
- (2) Do not require model fine-tuning;
- (3) Do not require online-optimization during inference;
- (4) Do not require paired source text of the reference image with word-to-word alignment;
- (5) Dispense with cumbersome attention modulations inside the denoising network;
- (6) Invariant to the specific architecture of the backbone diffusion model.

4 STUDY ON THE IMPACT OF λ

The coefficient λ in our method determines the proportion of the calibration phase to the non-calibration phase in the sampling trajectory. The impact of λ on the structure consistency, visual quality, and text fidelity of the generated images are shown in Fig. 1. When increasing the value of λ , the proportion of the calibration phase in the sampling trajectory is reduced, the degree of correlation

Table 1: 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 [8]	✓	✓	✗	✗	✗	✗
PAP [13]	✓	✓	✓	✗	✗	✗
Pix2Pix-zero [9]	✓	✓	✗	✗	✗	✗
InsPix2Pix [1]	✗	✓	✓	✓	✓	✓
PT-inversion [3]	✓	✓	✗	✗	✗	✗
StyleDiffusion [7]	✓	✓	✗	✗	✗	✗
VQCLIP [2]	✓	✓	✗	✓	✓	✗
DiffuseIT [6]	✓	✓	✗	✓	✓	✗
DiffusionCLIP [5]	✓	✗	✓	✓	✓	✗
Design Booster [11]	✗	✓	✓	✓	✓	✗
SINE [14]	✓	✗	✗	✗	✓	✓
Imagic [4]	✓	✗	✗	✓	✓	✓
FBSDiff (Ours)	✓	✓	✓	✓	✓	✓

**Figure 1: Study of the impact of λ to the structure consistency, visual quality, and text fidelity of the generated images.****Table 2: Quantitative evaluations of the text-driven I2I translation methods.**

Emphasis	Pursuing image appearance consistency					Pursuing image appearance divergence			
	Structure Similarity(↑)	LPIPS(↓)	AdaIN Style Loss(↓)	CLIP Similarity(↑)	Aesthetic Score(↑)	Structure Similarity(↑)	AdaIN Style Loss(↑)	CLIP Similarity(↑)	Aesthetic Score(↑)
PAP [13]	0.954	0.278	20.525	0.316	6.583	0.957	27.848	0.306	6.439
Null-text [8]	0.950	0.247	17.627	0.310	6.514	0.952	25.667	0.293	6.325
Pix2Pix-zero [9]	0.952	0.242	16.745	0.308	6.490	0.955	25.152	0.295	6.287
InsPix2Pix [1]	0.959	0.265	25.796	0.312	6.266	0.960	31.660	0.286	6.195
PT-inversion [3]	0.946	0.249	22.926	0.313	6.481	0.951	29.245	0.292	6.269
StyleDiffusion [7]	0.945	0.251	24.667	0.311	6.497	0.944	26.585	0.290	6.255
FBSDiff (ours)	0.962	0.240	15.302	0.314	6.566	0.958	34.725	0.309	6.464

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

between the reference image and the generated image is correspondingly weakened. Therefore, the Structure Similarity gradually decreases with the growth of λ . Besides, the increase of the calibration phase leads to more constrained I2I translation, which impairs model's ability to produce results that faithfully reflect the target text. Thus, the CLIP Similarity increases with the growth of λ . Moreover, we use NIMA (Neural Image Assessment) [12] to evaluate the quality of the translated images. Results show that the

image quality improves as λ increases. This is because increasing the value of λ corresponds to elongating the non-calibration phase, and thus allows to take greater advantage of the generative prior of the backbone diffusion model for image quality improvement. The impacts of λ on the above metrics are consistent for both the task pursuing image appearance consistency and the task favoring image appearance divergence. We empirically observe that

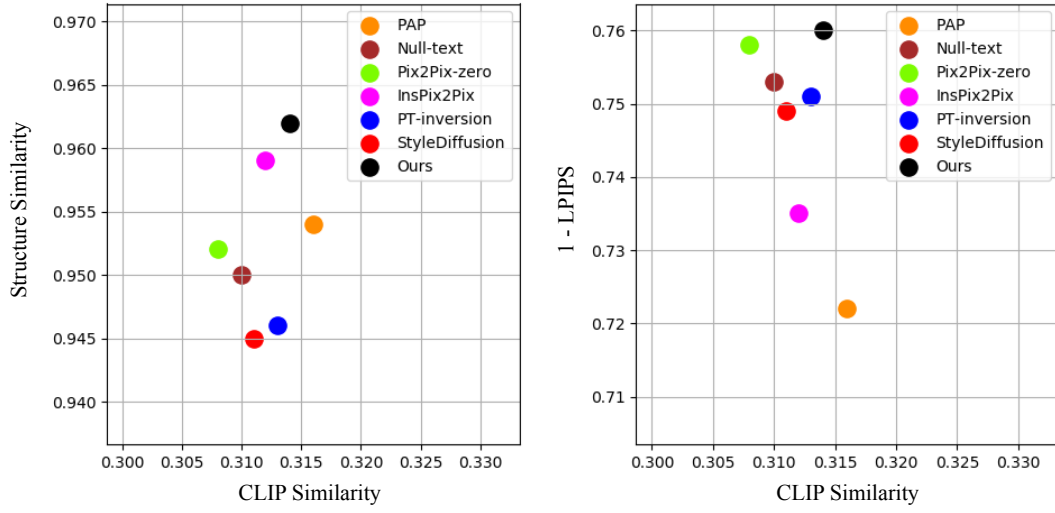


Figure 2: Visualization of method comparison for the image translation task pursuing image appearance consistency. Left: comparison in CLIP Similarity (↑) and Structure Similarity (↑). Right: comparison in CLIP Similarity (↑) and 1-LPIPS (↑). Our method with low-FBS achieves the most top-right position in both two scatters.

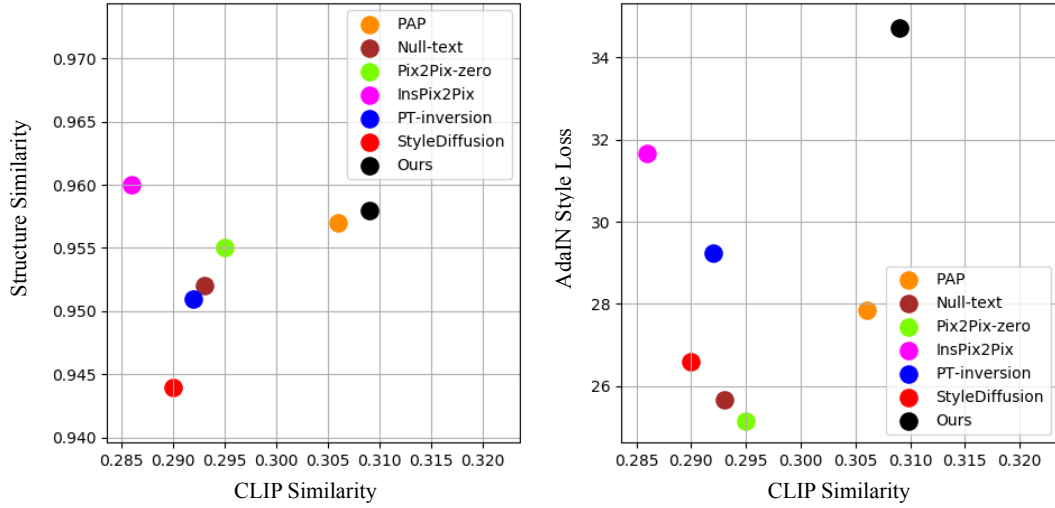


Figure 3: Visualization of method comparison for the image translation task pursuing image appearance divergence. Left: comparison in CLIP Similarity (↑) and Structure Similarity (↑). Right: comparison in CLIP Similarity (↑) and AdaIN Style Loss (↑). Our method with high-FBS achieves the most top-right position in both two scatters.

$\lambda \in (4, 5)$ provides better trade-off among text fidelity, structure consistency, and visual quality.

4.1 Quantitative Results Visualization

For the quantitative evaluation reported in Tab. 2, we visualize partial results to highlight the superiority of our method over related approaches. For the I2I task pursuing image appearance consistency, we visualize the scatter plot about Structure Similarity (↑) and CLIP Similarity (↑), and the scatter plot about (1-LPIPS) (↑) and

CLIP Similarity (↑). Results in Fig. 2 show that our method achieves the most top-right position in both two scatter plots, indicating the overall superiority of our method with low-FBS in text fidelity and appearance consistency. For the I2I task pursuing image appearance divergence, we visualize the scatter plot about Structure Similarity (↑) and CLIP Similarity (↑), and the scatter plot about AdaIN Style Loss (↑) and CLIP Similarity (↑). Results in Fig. 3 show the most top-right position achieved by our method in both two plots, which reflects the outstanding ability of our method with high-FBS to

produce diverse image appearance while guaranteeing structure consistency and text fidelity. At last, we showcase more examples in Fig. 30 about controllable guiding intensity of the reference image by modulating the bandwidth of the substituted frequency band in low-FBS.

5 MORE RESULTS

Below we showcase more text-driven I2I translation results of our FBSDiff. Results of our method with low-FBS for image appearance and layout control are displayed in Fig. 4 to Fig. 13. Results of our method with high-FBS for image contour control are displayed in Fig. 14 to Fig. 23. Our results with mid-FBS for image layout control are displayed in Fig. 24 to Fig. 28. In Fig. 29, we showcase more I2I results of our method realized by low-FBS, mid-FBS, and high-FBS.

REFERENCES

- [1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.
- [2] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*. Springer, 88–105.
- [3] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. 2023. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7430–7440.
- [4] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.
- [5] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2426–2435.
- [6] Gihyun Kwon and Jong Chul Ye. 2022. Diffusion-based Image Translation using disentangled style and content representation. In *The Eleventh International Conference on Learning Representations*.
- [7] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. 2023. Stylediffusion: Prompt-embedding inversion for text-based editing. *arXiv preprint arXiv:2303.15649* (2023).
- [8] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6038–6047.
- [9] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. 2023. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.
- [10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.
- [11] Shiqi Sun, Shancheng Fang, Qian He, and Wei Liu. 2023. Design Booster: A Text-Guided Diffusion Model for Image Translation with Spatial Layout Preservation. *arXiv preprint arXiv:2302.02284* (2023).
- [12] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural image assessment. *IEEE transactions on image processing* 27, 8 (2018), 3998–4011.
- [13] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1921–1930.
- [14] Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N Metaxas, and Jian Ren. 2023. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6027–6037.



Figure 4: More results of our method with low-FBS for appearance and layout control.

I2I translation results with low-FBS (Example 2)

Reference



"a boy"



"young lady"



"elderly man"



"soldier"



Figure 5: More results of our method with low-FBS for appearance and layout control.

I2I translation results with low-FBS (Example 3)

Reference



"young lady"



"elderly woman"



"little boy"



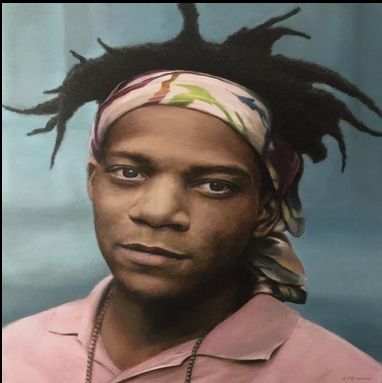
"elderly man"



Figure 6: More results of our method with low-FBS for appearance and layout control.

I2I translation results with low-FBS (Example 4)

Reference



"little boy"



"young lady"



"elderly lady"



"robot"



Figure 7: More results of our method with low-FBS for appearance and layout control.

I2I translation results with low-FBS (Example 5)

Reference



"castle"



"ancient ruins"



"modern house"



"Chinese ancient buildings"



Figure 8: More results of our method with low-FBS for appearance and layout control.

I2I translation results with low-FBS (Example 6)

Reference



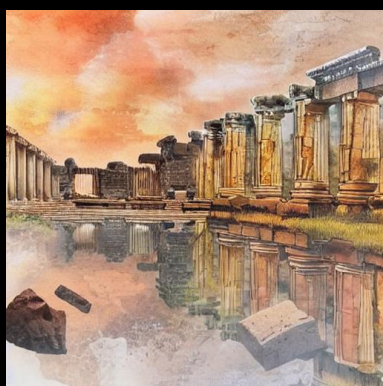
"city street"



"royal palace"



"ancient ruins"



"train"



Figure 9: More results of our method with low-FBS for appearance and layout control.

I2I translation results with low-FBS (Example 7)

Reference



"castle"



"city street"



"temple"



"train station"



Figure 10: More results of our method with low-FBS for appearance and layout control.

I2I translation results with low-FBS (Example 8)

Reference



"castle"



"city street"



"factories"



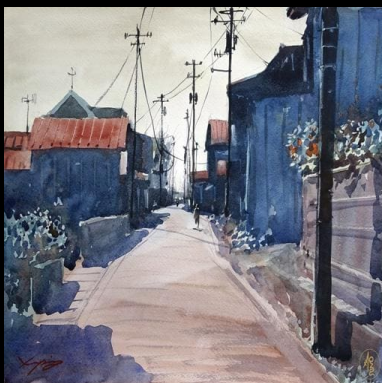
"amusement park"



Figure 11: More results of our method with low-FBS for appearance and layout control.

I2I translation results with low-FBS (Example 9)

Reference



"car race"



"train"



"modern city"



"mountain landscape"



Figure 12: More results of our method with low-FBS for appearance and layout control.

I2I translation results with low-FBS (Example 10)

Reference



"park"



"ruins"



"royal palace"



"ancient Rome"



Figure 13: More results of our method with low-FBS for appearance and layout control.

I2I translation results with high-FBS (Example 1)

Reference



"water color"



"oil painting"



"pencil sketch"



"wooden sculpture"



Figure 14: More results of our method with high-FBS for contour control.

I2I translation results with high-FBS (Example 2)

Reference



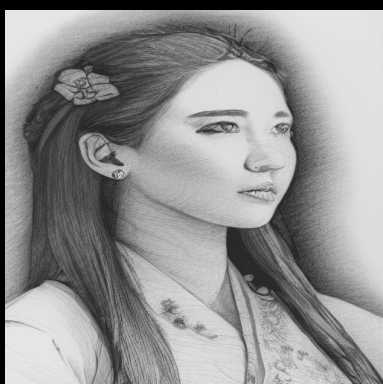
"water color"



"anime figure"



"pencil sketch"



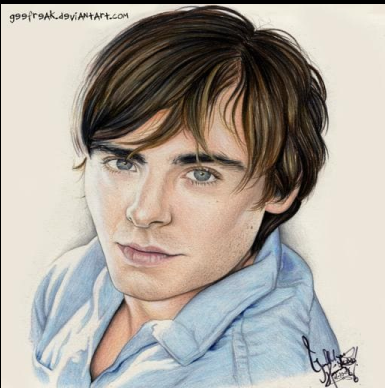
"sculpture"



Figure 15: More results of our method with high-FBS for contour control.

I2I translation results with high-FBS (Example 3)

Reference



"robot"



"plaster model"



"pencil sketch"



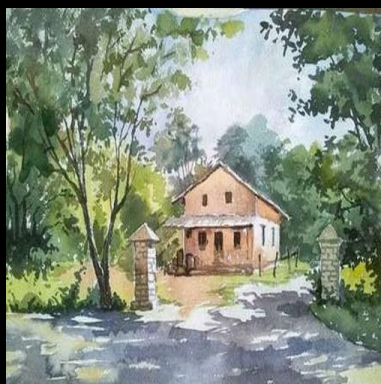
"wooden sculpture"



Figure 16: More results of our method with high-FBS for contour control.

I2I translation results with high-FBS (Example 4)

Reference



"autumn"



"winter"



"crayon drawing"



"ink-wash painting"

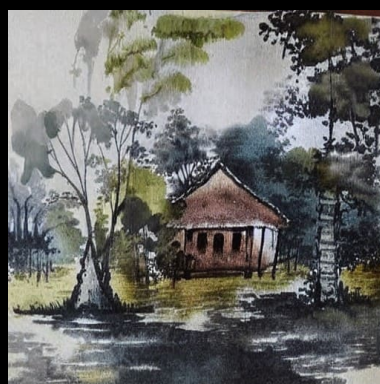


Figure 17: More results of our method with high-FBS for contour control.

I2I translation results with high-FBS (Example 5)

Reference



"autumn"



"winter"



"crayon drawing"



"water color"



Figure 18: More results of our method with high-FBS for contour control.

I2I translation results with high-FBS (Example 6)

Reference



"water color"



"iceberg"



"colored pencil sketch" "desert landscape"



Figure 19: More results of our method with high-FBS for contour control.

I2I translation results with high-FBS (Example 7)

Reference



"autumn"



"winter"



"oil painting"



"Chinese landscape painting"



Figure 20: More results of our method with high-FBS for contour control.

I2I translation results with high-FBS (Example 8)

Reference



"autumn"



"winter"



"oil painting"



"crayon drawing"



Figure 21: More results of our method with high-FBS for contour control.



Figure 22: More results of our method with high-FBS for contour control.

I2I translation results with high-FBS (Example 10)

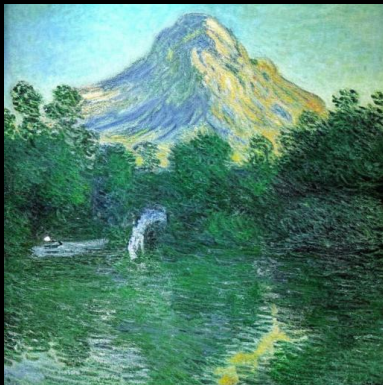
Reference



"desert"



"Monet painting"



"Van Gogh painting" "Chinese ink-wash painting"

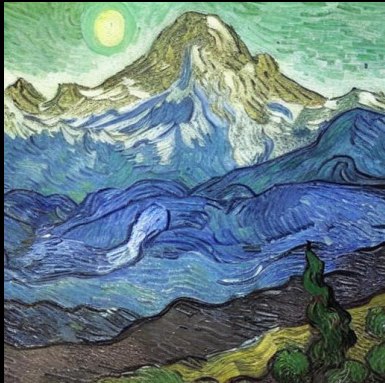


Figure 23: More results of our method with high-FBS for contour control.

I2I translation results with mid-FBS (Example 1)

Reference



"science lab"



"bed room"



"shopping mall"



"gym"



Figure 24: More results of our method with mid-FBS for layout control.

I2I translation results with mid-FBS (Example 2)

Reference



"royal palace"



"painting of a garden"



"Rome ruins"



"space station"

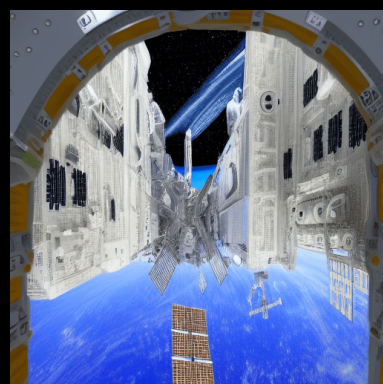


Figure 25: More results of our method with mid-FBS for layout control.

I2I translation results with mid-FBS (Example 3)

Reference



"LEGO brick"



"space center"



"Birthday cake"



"snow man"



Figure 26: More results of our method with mid-FBS for layout control.

I2I translation results with mid-FBS (Example 4)

Reference



"train"



"museum"



"aquarium"



"science lab"



Figure 27: More results of our method with mid-FBS for layout control.

I2I translation results with mid-FBS (Example 5)

Reference



"ruins"



"lake"



"railway track"



"science lab"

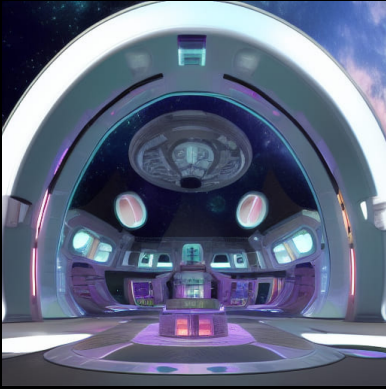


Figure 28: More results of our method with mid-FBS for layout control.



Figure 29: More visual comparison among the I2I results realized by low-FBS, mid-FBS, and high-FBS.

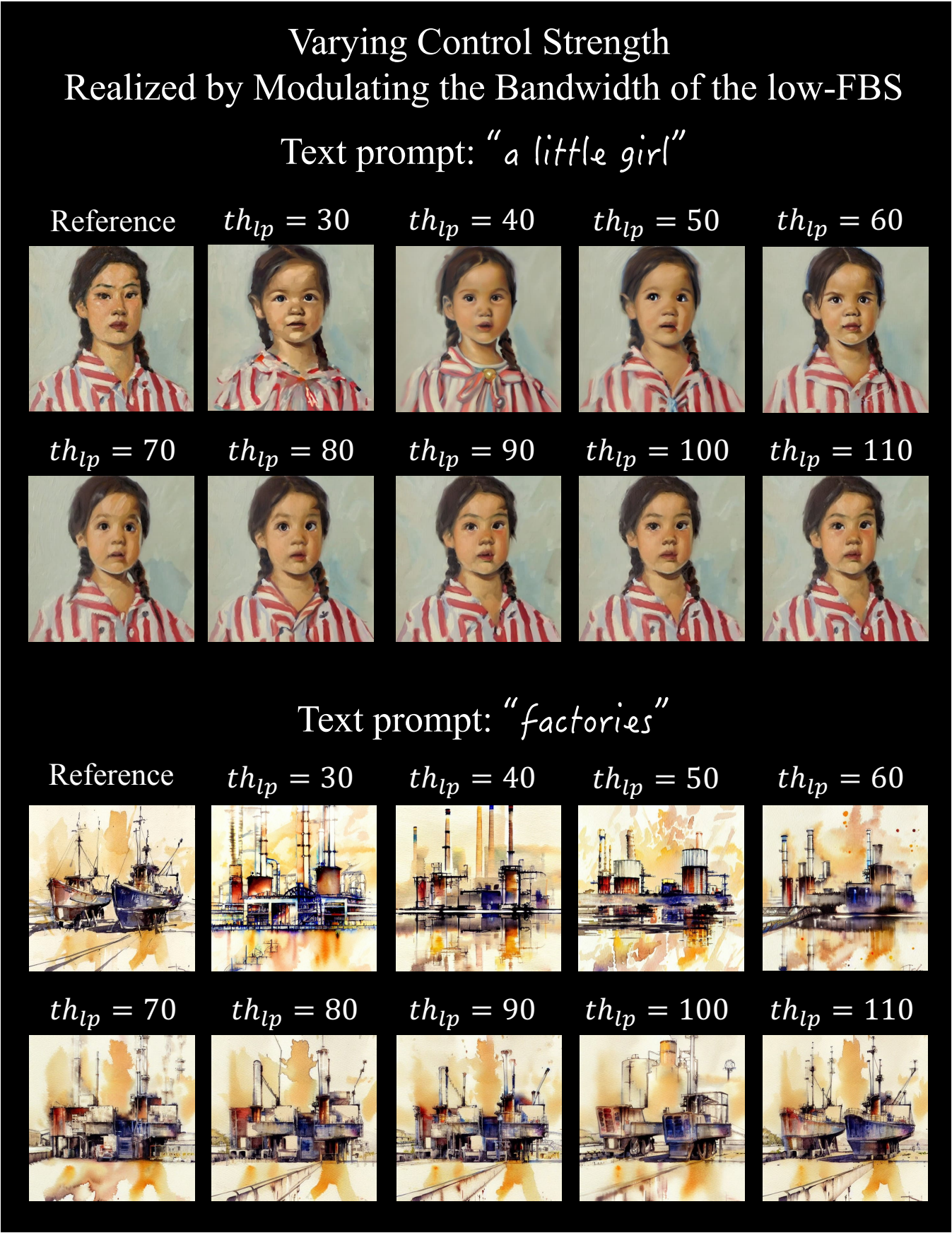


Figure 30: More visual comparison among the I2I results realized by low-FBS, mid-FBS, and high-FBS.