Supplementary Material for Frequency-Guided Diffusion for Training-Free Text-Driven Image Translation

Zheng Gao[1,2*], Jifei Song[2], Zhensong Zhang[2], Jiankang Deng[3], Ioannis Patras[1]
[1]Queen Mary University of London, [2]Huawei London Research Center, [3]Imperial College London

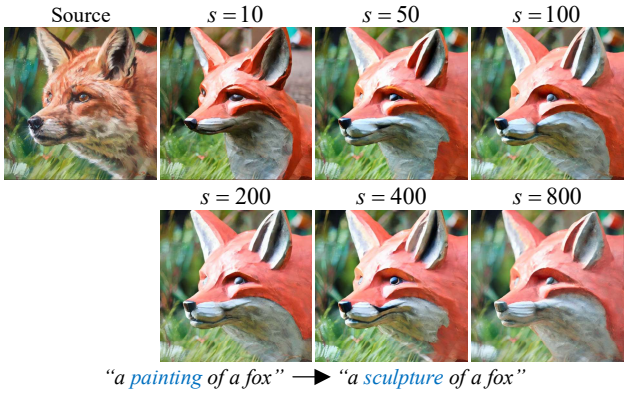


Figure 1. **Ablation on high-frequency alignment scale** s.

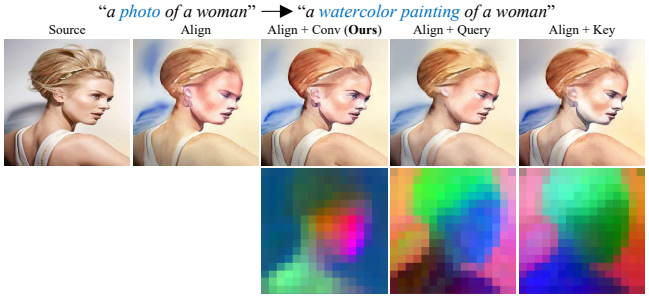
1. Additional ablation study

1.1. Effect of high-frequency alignment scale

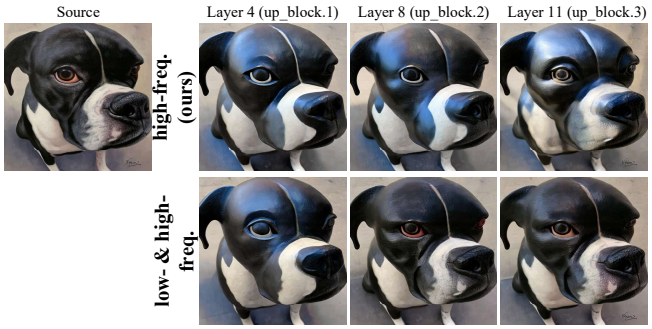
In Fig. 1, we study the effect of high-frequency scale s. As s increases, the high-frequency guidance starts to take effect by aligning structure with source (*e.g.*, mouth). Further increasing could result in large gradient and affect structure preservation and image quality (*e.g.*, blurred generation in the rightmost column). We empirically find that it works well within the range of $[100, 200]$. These discussions are also observed in Tab. 1a, suggesting our high-frequency guidance is robust to the choice of scale.

1.2. Effect of feature and layer choice

In Fig. 2, we ablate the choice of feature and layer in high-frequency injection. Since prior work [19] observes that decoder features provide better cues for structure preservation than encoder, we use the decoder features for high-frequency injection. In Fig. 2a, the comparison of generated images between convolution and self-attention features shows that convolution features are preferable for structure preservation, which is aligned with the observation in [19]. The quantitative results in Tab. 1b also indicate the convolution feature is a better choice for structure preservation. In Fig. 2b, we compare the convolution features from different decoder blocks. Along with our high-frequency alignment, the high-frequency injection at commonly-used layer 4 is enough for structure preservation. Deeper layers with higher resolutions (*e.g.*, layer 11) incorporates more appearance information and form the final prediction. This might cause appearance leakage (*e.g.*, eyes and mouth), which is consistent with the observations in [19]. Note that as we



(a) Feature choice. **Align** denotes high-frequency alignment. **Conv** denotes high-frequency injection with convolution features. **Query/Key** denotes high-frequency injection with self-attention features query/key, respectively. PCA visualization of convolution and attention features are provided below the image, showing that convolution features have better disentanglement of object parts (*e.g.*, face and hair). Therefore, convolution features are more suitable for compensating our high-frequency alignment.



(b) Layer choice. The convolution features from decoder upsample block up_block.[1,2,3] are used. First row only injects high-frequency components of features (*i.e.*, ours) while second row injects both low- and high-frequency components (*i.e.*, direct replacement of features used by prior works).

Figure 2. **Ablation on feature and layer choice in high-frequency injection**.

s	DINO ↓	Edge ↑	CLIP ↑	HPS ↑
0	0.037	0.915	0.270	**27.76**
100	0.034	0.930	0.271	27.64
200	0.035	**0.931**	**0.271**	27.56
800	**0.031**	0.928	0.269	27.25

(a) High-frequency alignment scale s.

Variant	DINO ↓	Edge ↑	CLIP ↑	HPS ↑
Query	0.042	0.921	0.265	27.14
Key	0.038	0.929	**0.272**	27.48
Conv (**Ours**)	**0.035**	**0.931**	0.271	**27.56**

(b) Feature choice in high-frequency injection.

r_{lp}	CSD ↑	CLIP ↑	HPS ↑
0	0.554	0.259	27.27
3	0.637	**0.262**	**27.38**
5	0.649	0.260	27.34
20	**0.701**	0.249	26.91

(c) Threshold r_{lp} for low-frequency guidance.

r_{hp}	r_{hp}^{pix}	DINO ↓	Edge ↑	CLIP ↑	HPS ↑
0	20	**0.031**	**0.935**	0.266	27.61
3	10	0.033	0.933	**0.273**	27.54
3	20	0.035	0.931	0.271	27.56
3	30	0.036	0.927	0.271	27.62
3	50	0.038	0.913	0.270	**27.70**
5	20	0.038	0.929	0.270	27.65
10	20	0.041	0.926	0.269	27.68

(d) Threshold r_{hp}^{pix}/r_{hp} for high-frequency guidance.

Table 1. **Quantitative ablations on Aesthetics [18]**. The default setting is highlighted.

only inject high-frequency components of convolution features, the color leakage is less apparent but there are still traces at layer 11. Therefore, we use convolution features from layer 4 of decoder for high-frequency injection.

1.3. Quantitative ablations on frequency threshold

In addition to the qualitative ablation on frequency threshold in main paper, we provide quantitative results in Tabs. 1c and 1d. For low-frequency guidance threshold (Tab. 1c), extremely large r_{lp} (*e.g.*, 20) may have better style preservation but fails to follow target prompt. We empirically find that $r_{lp} = 5$ has better balance between style preservation and image-text alignment. As for high-frequency guidance threshold (Tab. 1d), we find that it works well for $r_{hp} \leq 5$ and $r_{hp}^{pix} \leq 30$. In practice, as discussed in main paper, we choose $r_{hp} = 3$ as it has better balance between fidelity to source image's content and image-text alignment. For r_{hp}^{pix}, we choose $r_{hp}^{pix} = 20$ as it has better balance between structure preservation and image-text alignment.

Method	Style-guided			Structure-guided			
	CSD ↑	CLIP ↑	HPS ↑	DINO ↓	Edge ↑	CLIP ↑	HPS ↑
Image editing method							
CycleDiff [20]	0.601	0.270	26.67	0.054	0.900	0.258	27.04
DDPM-Edit [8]	0.644	0.271	26.85	0.038	0.925	0.265	27.51
LEDITS++ [1]	0.715	0.254	26.70	0.031	0.937	0.240	27.13
FlexiEdit [10]	0.663	0.260	26.11	0.094	0.882	0.230	25.89
Image translation method							
P2P+NT [5, 15]	**0.662**	0.244	26.14	**0.025**	**0.942**	0.238	26.66
PnP [19]	0.543	0.265	26.39	0.052	0.915	0.259	26.94
P2P-Zero [16]	0.617	0.243	25.94	0.059	0.886	0.218	24.37
FreeControl [14]	0.512	0.270	**27.01**	0.063	0.907	0.267	**27.45**
Ctrl-X [13]	0.354	**0.271**	26.60	0.144	0.843	**0.275**	27.09
PIC [11]	0.507	0.251	25.96	0.049	0.903	0.243	26.56
FCDDiff [4]†	0.538	0.252	26.30	0.036	0.928	0.252	26.67
FBSDiff [3]	0.610	0.260	26.71	0.053	0.909	0.264	27.16
FGD (**Ours**)	**0.718**	**0.274**	**26.93**	**0.032**	**0.935**	**0.269**	**27.57**

Table 2. **Additional quantitative results on ImageNet-R-TI2I [19]**. †: training-based method.

2. Additional experimental results

2.1. Additional quantitative results

Following [8, 19], in Tab. 2, we provide style- and structure-guided translation results on ImageNet-R-TI2I (IN-R) [19]. Our method achieves competitive results on both tasks. Note that although LEDITS++ [1] and P2P+NT [5, 15] have strong style/structure preservation performance, they have low CLIP score, suggesting these works struggle with following target prompt instruction to change the source image. FreeControl [14] and Ctrl-X [13] are able to generate high-quality images (high HPS score) with strong image-text alignment (high CLIP score). However, they are inferior with style preservation (low CSD score) and structure preservation (high DINO distance, low Edge similarity), which is also observed in Fig. 4 of main paper and in Figs. 4 and 5 of supplementary.

2.2. Adaption to Stable Diffusion v1.5

By default, we use Stable Diffusion v2.1 (SDv2.1) [17] as pre-trained T2I model. In Tab. 3, we apply our method to Stable Diffusion v1.5 (SDv1.5) and show that it generalizes well on models with different capacities. In Fig. 3, we provide qualitative results and show that compared with SDv1.5 variant, SDv2.1 variant is able to generate images with better quality (*e.g.*, facial details in 4$^{\text{th}}$ row), more details (*e.g.*, hair in 2$^{\text{nd}}$ row) and better image-text alignment (*e.g.*, watercolor style in 5$^{\text{th}}$ row). However, despite the fact that SDv1.5 variant is less expressive, it still successfully follows the target prompt and preserves the essential

Method	Backbone	Style-guided			Structure-guided			
		CSD ↑	CLIP ↑	HPS ↑	DINO ↓	Edge ↑	CLIP ↑	HPS ↑
DDPM-Edit [8]	SDv1.5	0.645	0.272	26.67	0.041	0.922	0.265	27.21
FBSDiff [3]	SDv1.5	0.587	0.270	26.59	0.052	0.910	0.265	27.02
Ours	SDv1.5	0.690	0.273	26.70	0.034	0.931	0.267	27.30
DDPM-Edit [8]	SDv2.1	0.644	0.271	26.85	0.038	0.925	0.265	27.51
FBSDiff [3]	SDv2.1	0.610	0.260	26.71	0.053	0.909	0.264	27.16
Ours	SDv2.1	0.718	0.274	26.93	0.032	0.935	0.269	27.57

Table 3. **Comparison of results with SDv1.5 and SDv2.1 on ImageNet-R-TI2I [19].**

Method	Style-guided					Structure-guided					
	CSD ↑	CLIP ↑	HPS ↑	FID ↓	Time ↓	DINO ↓	Edge ↑	CLIP ↑	HPS ↑	FID ↓	Time ↓
P2P+NT [5, 15]	**0.669**	0.236	26.41	136.8	144s	**0.027**	**0.950**	0.252	26.98	112.3	144s
PnP [19]	0.477	0.254	26.63	192.6	77s	0.044	0.919	**0.270**	26.99	192.6	77s
P2P-Zero [16]	0.426	0.237	26.09	219.9	76s	0.046	0.920	0.230	26.29	142.0	76s
FreeControl [14]	0.436	0.252	**27.22**	227.3	182s	0.045	0.913	0.270	**27.42**	179.1	182s
Ctrl-X [13]	0.230	**0.264**	26.94	263.8	17s	0.123	0.860	0.269	26.16	228.7	17s
PIC [11]	0.421	0.254	26.65	222.3	17s	0.053	0.909	0.243	26.32	184.1	17s
FCDDiff [4]†	0.434	**0.261**	26.39	239.62	8s	0.038	0.928	0.231	26.36	173.01	8s
FBSDiff [3]	0.436	0.257	26.38	253.1	86s	0.058	0.908	0.253	26.58	166.6	86s
FGD (**Ours**)	**0.649**	**0.260**	**27.34**	151.7	12s	**0.035**	**0.931**	**0.271**	**27.56**	160.4	25s

Table 4. **Comparison of sampling time on LAION-Aesthetics 6.5+ [18].** The running time on one NVIDIA 3090 GPU is reported. We additionally report FID (Fréchet Inception Distance) [6]. However, because the translated image may not have the same distribution as the original input image, most baselines in the area of image translation don't use FID.

style/structure in the generation.

2.3. Additional qualitative results

In Figs. 4 and 5, we provide more qualitative results. As observed in main paper, compared with baselines, our method is able to generate images that preserve source image's style/structure and match target text prompt. In particular, for style-guided translation, our method has better detail preservation (*e.g.*, clothes and hair in 3rd row of Fig. 4) and color alignment with source image (*e.g.*, 4th row of Fig. 4). For structure-guided translation, our method has better ability of preserving edges and contour (*e.g.*, facial details in 2nd row, big and small boats in 4th row of Fig. 5).

2.4. Diverse sampling results

In Fig. 6, we show that our FGD is able to generate diverse sampling results for the fixed target prompt by applying DDPM inversion [8] to invert initial latent through adding Gaussian noise.

3. Sampling time

In Tab. 4, we compare the sampling time of training-free image translation methods. Our method achieves competitive results against prior works with relatively low cost. In particular, on style-guided translation, our time is much lower than prior works as we don't rely on long step inversion (*e.g.*, 50 *vs*. 1000 steps used by FBSDiff [3]). Although P2P+NT [5, 15] exhibits strong style and structure preservation performance, as discussed in main paper, it often fails to follow target prompt and leaves the source image unchanged, which is indicated by the low CLIP score (image-text alignment).

4. Additional implementation details

4.1. Implementation

We use layer 4 of decoder for high-frequency injection as prior work [19] observes that compared with deeper layers with higher resolution, it contains structure information without resulting in appearance leakage from source image. The spatial size of convolution features is 16×16. Following common practice [3, 19], CFG [7] scale is 7.5 and DDIM sampling step is 50 with $\eta = 1$.

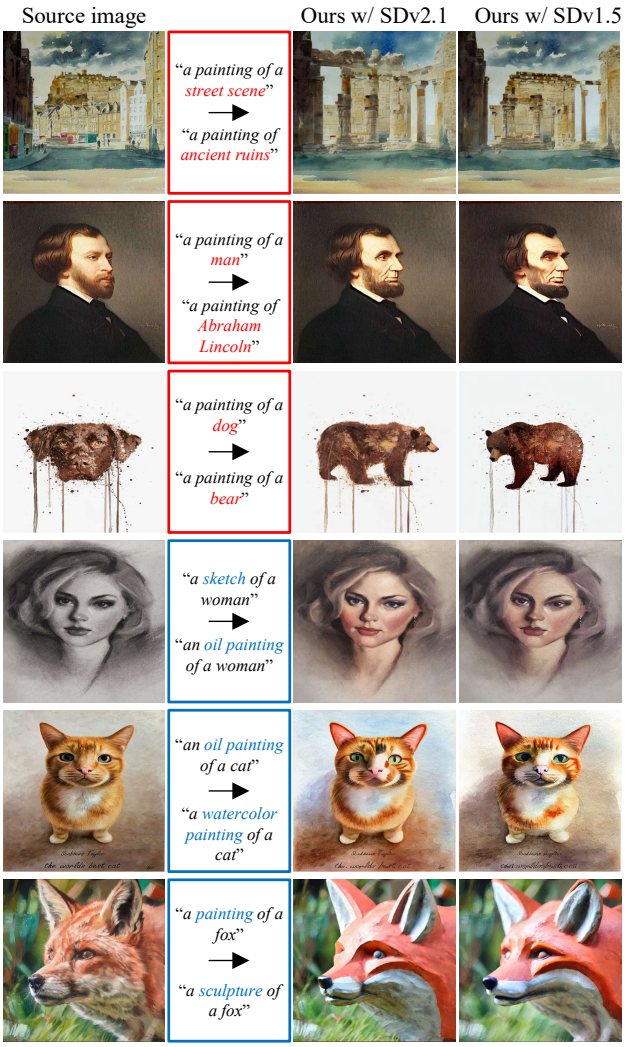


Figure 3. **Qualitative results with SDv1.5 and SDv2.1**.

4.2. Dataset

LAION-Aesthetics 6.5+ [18]. Following FBSDiff [3], we evaluate on LAION-Aesthetics 6.5+ [18], a subset of LAION-5B [18] containing high-quality images with aesthetic scores of 6.5 or higher. As the selected samples and adopted prompts used in the evaluation of FBSDiff [3] are not available, we create a benchmark by sampling 200 source images from Aesthetics, 100 for each task of style-/structure-guided translation. These images include paintings/photos of humans, animals and other objects (*e.g.*, boat, flower, plane). For style-guided translation, 80% are paintings while the others are photos. For structure-guided translation, 60% are paintings while the others are photos. Samples of these images are provided in Fig. 7. We first use BLIP [12] to automatically generate source text prompts describing these images. Then we manually revisit the generated source prompts to ensure they accurately describe the source images. For each task of style-/structure-guided translation, we manually create 3 target prompts for each pair of source image and source prompt. For style-guided translation, we replace the object in source prompt by using target prompt to specify a different object class, *i.e.*, class that is semantically related to the class in source prompt (*e.g.*, "*a painting of a man*" → "*a painting of a boy*"). For structure-guided translation, we use target prompt to specify a different style (*e.g.*, "*a photo of a dog*" → "*an oil painting of a dog*"). The new style is randomly sampled from a list of style choices: oil painting, watercolor painting, pencil sketch, sculpture and origami.

PIE-Bench [9]. PIE-Bench [9] is a prompt-driven image editing benchmark. We use subset "*1 Change Object*" (*i.e.*, change an object to another) for style-guided translation and subset "*9 Change Style*" (*i.e.*, change the image style) for structure-guided translation. Each of the subset has 80 images, including paintings and photos of humans, animals and scenes.

ImageNet-R-TI2I [19]. ImageNet-R-TI2I [19] (IN-R) is a widely-used dataset in image translation and editing [8, 14, 19], with various renditions of 10 classes from ImageNet [2]. Each image has 5 target prompts. Among them, 3 prompts are structure-guided text instructions (*e.g.*, "*a sketch of a penguin*" → "*a toy of a penguin*") while the other 2 prompts modify both style and structure (*e.g.*, "*a sketch of a penguin*" → "*a sculpture of a swan*"). We directly use the 3 structure-guided instructions for our structure-guided translation task. For style-guided translation, we keep the style unchanged in the prompt (*e.g.*, "*a sketch of a penguin*" → "*a sketch of a swan*").

References

[1] Manuel Brack, Felix Friedrich, Katharina Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8861–8870, 2024. 2

[2] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4

[3] Xiang Gao and Jiaying Liu. Fbsdiff: Plug-and-play frequency band substitution of diffusion features for highly controllable text-driven image translation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 4101–4109, New York, NY, USA, 2024. Association for Computing Machinery. 2, 3, 4

[4] Xiang Gao, Zhengbo Xu, Junhan Zhao, and Jiaying Liu. Frequency-controlled diffusion model for versatile text-guided image-to-image translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3):1824–1832, 2024. 2, 3

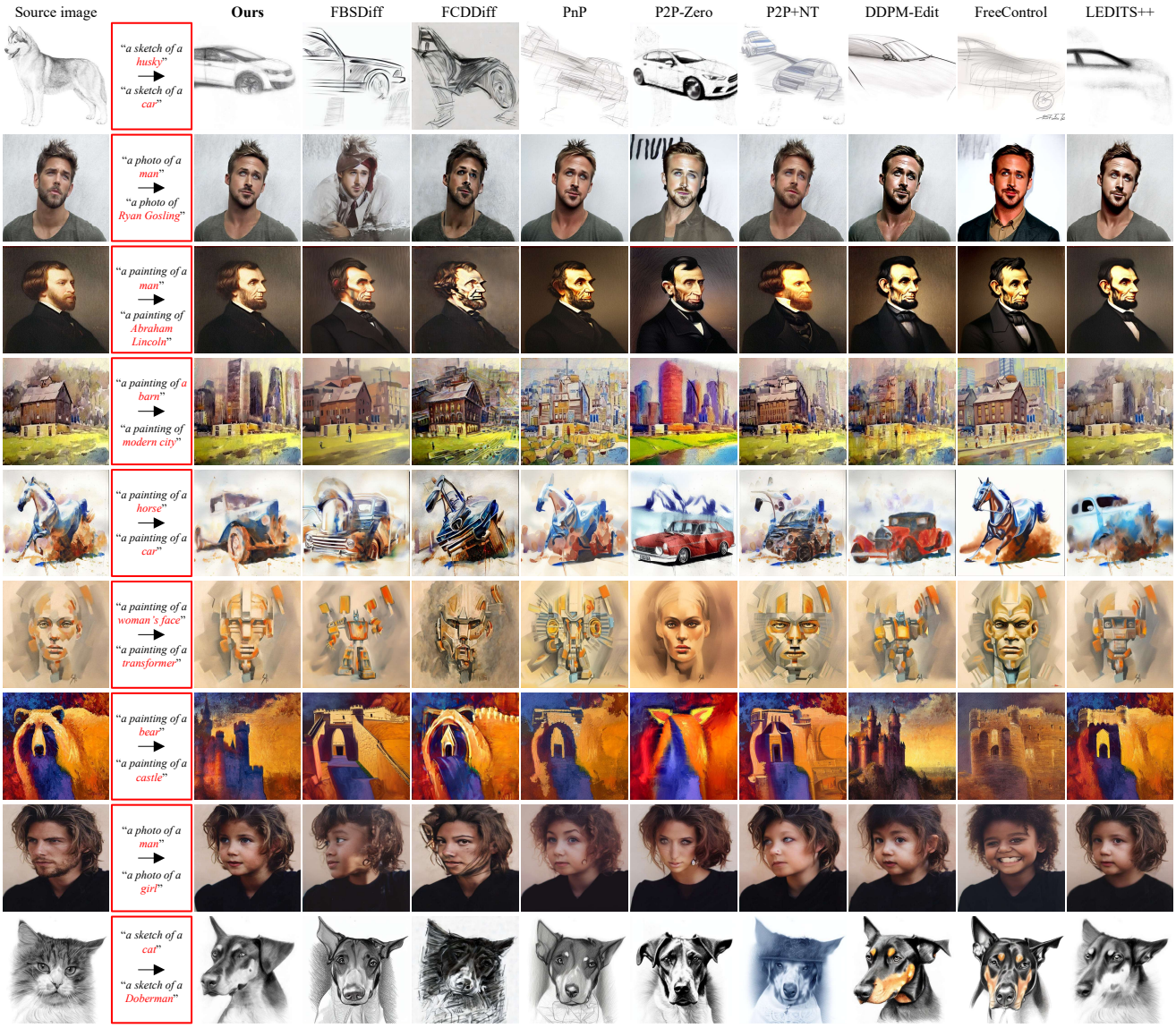


Figure 4. **Additional qualitative results on style-guided translation** (Low-frequency guidance for frequency-based models).

[5] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 3

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3

[8] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12469–12478, 2024. 2, 3, 4

[9] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *The Twelfth International Conference on Learning Representations*, 2024. 4

[10] Gwanhyeong Koo, Sunjae Yoon, Ji Woo Hong, and Chang D Yoo. Flexiedit: Frequency-aware latent refinement for enhanced non-rigid editing. In *Computer Vision – ECCV 2024*. Springer Nature Switzerland, 2024. 2

[11] Junsung Lee, Minsoo Kang, and Bohyung Han. Diffusion-based image-to-image translation by noise correction via prompt interpolation. In *Computer Vision – ECCV 2024*. Springer Nature Switzerland, 2024. 2, 3

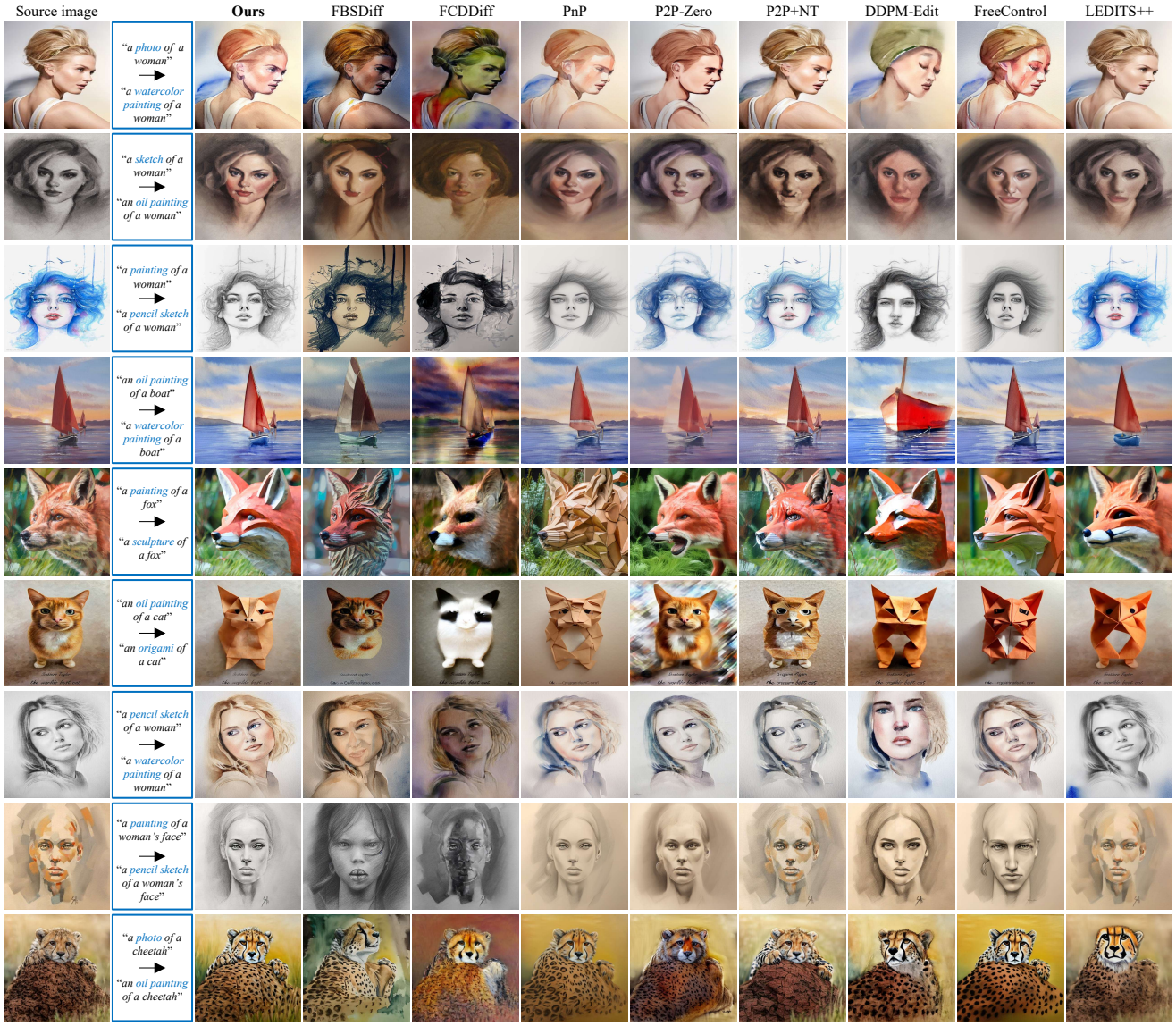


Figure 5. **Additional qualitative results on structure-guided translation** (High-frequency guidance for frequency-based models).

[12] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 4

[13] Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. Ctrl-x: Controlling structure and appearance for text-to-image generation without guidance. In *Advances in Neural Information Processing Systems*, 2024. 2, 3

[14] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7465–7475, 2024. 2, 3, 4

[15] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6038–6047, 2023. 2, 3

[16] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, New York, NY, USA, 2023. Association for Computing Machinery. 2, 3

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*

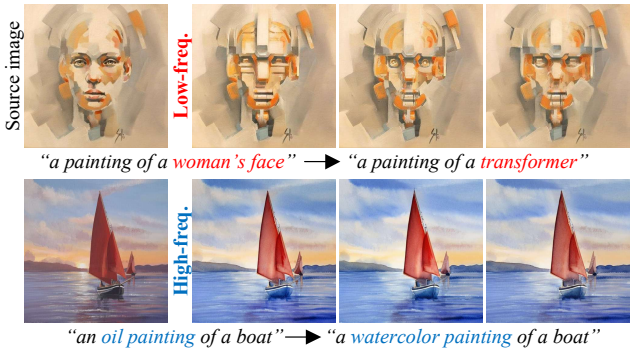


Figure 6. **Diverse sampling results with fixed target prompt**. Top: eyes are different. Bottom: sunset is different.

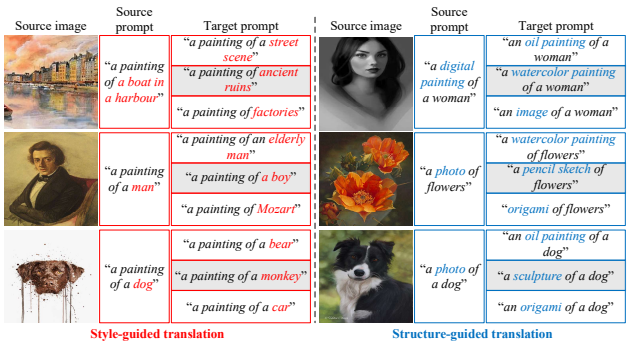


Figure 7. **Samples from Aesthetics [18]**.

(CVPR), pages 10674–10685, 2022. 2

[18] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2, 3, 4, 7

[19] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1921–1930, 2023. 1, 2, 3, 4

[20] Chen Henry Wu and Fernando De La Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7344–7353, 2023. 2