Supplementary Materials: FreePIH: Training-Free Painterly Image Harmonization with Diffusion Model

Anonymous Authors

1 IMPLEMENTATION DETAILS

The foreground items in our experiment are extracted from COCO datasets[4], which is a large-scale object detection, segmentation, and captioning dataset. COCO has 80 object categories, 1.5 million object instances, we use the pycocotools library to fetch content and mask from this dataset. The background images are randomly selected from LAION[9] (acronym for Large-scale Artificial Intelligence Open Network) which is a number of large datasets of image-caption pairs. Now the dataset contain more than 5 billion image-text pairs of various artistic styles. For all the baselines, we download the code from their their official gitHub repositories.

For our FreePIH project, we utilize the pre-trained Latent Diffusion Model(LDM) developed by Stability AI [8]. The LDM consists of several key components: an Image Encoder, a Text Encoder, an Image Decoder, and a DGM module. The Image Encoder transforms the original image into a latent feature representation, significantly reducing memory usage during subsequent calculations. The Text Encoder directly utilizes the pre-trained CLIP text encoder. The DGM module follows a U-Net architecture, incorporating ResNet-Blocks, Spatial Transformers (Cross Attention), and DownSample/UpSample convolution layers.

Prior to inputting the latent feature into the DGM, the LDM employs a time embedding module to encode the noise level (t), which is then concatenated with the input latent feature. These augmented features are progressively processed through the DGM modules until the noise level (t) reaches 0. Subsequently, the Image Decoder receives the final denoised latent feature and decodes it into the original image space.

In our modified version of the LDM, we have introduced several additional loss terms to optimize the latent feature, with the goal of transferring its style. It is worth noting that the only learnable part in our pipeline is the latent feature of the foreground items. During the inference process, we freeze all the modules including the Image Encoder, Text Encoder, DGM, and Image Decoder. This approach allows us to avoid heavy training costs and enables quick utilization for painterly image harmonization sourced from the internet.

We implement FreePIH and test all the baselines on ubuntu 18.04 LTS operation system, with 64GB memory, a 12900K Intel CPU @3.20GHz and an NVIDIA RTX 4090 GPU. The pytorch version is 2.0.0. And the output image size is 512×512.

1.1 Details of questionnaire

Our questionnaire are as shown in Figure 1. Users were asked to vote for the top-1 harmonious image by question "Which of the following images works best for the fusion of the two iamges" and give their score at the same time by question "Give your rating for the fusion of each graph, with 1 (Bad) to 5(Excellent)."

- A red bus on the beach of an oil painting.
- A sunflower on the upper right of a sunflower oil painting with Leonid Afremov style.
- A painting of a pyramid in the countryside. The pyramid is near a lake. One boat is on the lake.
- A bus in the bottle left of an oil painting. There are many people and trees on the road.
- A yellow boat is on the river. The sun is going down.
- A painting of a Sphinx in the countryside. The Sphinx is near a house and river.
- A moai statue in an oil painting with various colors, the moai statue is near a tree and a street light.
- A snow scene, a bench against a wall, a man walking in the middle.

1.2 Details of Baselines

The details about these baselines are as follows:

- **PIE[7]**: Poisson image editing is the most classical paper about image editing which uses generic interpolation machinery based on solving Poisson equations for seamless editing of image regions.
- **DIB[10]**: Deep Image Blending is an updated version of Poisson image editing with the combination of several other loss terms.
- **SD-Text[8]**: Given a new text description about the composition image, use pretrain stable diffusion weight to generate a new image correspond to the text input.
- **SDEdit[6]**: Given the composition images and text input as guidance, Stochastic Differential Editing use pretrain stable diffusion model to iteratively refine the mask range.
- **PHDNet[2]**: A novel painterly harmonization network consisting of a dual-domain generator and a dual-domain discriminator, which harmonizes the composite image in both spatial domain and frequency domain.
- **BlendDM[1]**: A novel DM-based text driven image editing methods. Given an input image and a mask, BlendDM modifies the masked area according to a guiding text prompt, without affecting the unmasked regions.
- **CDC[3]**: CDC incorporates high-frequency background details and low-frequency foreground style for DM image generation.
- **InST[11]**: It uses CLIP image encoder to translate a style image into the text domain embedding, then use the embedding to guide the DM generation process.
- **PHDiff[5]**: It uses the trained lightweight adaptive encoder and dual encoder fusion to guide the DM denoising process of diffusion.

2 ADDITIONAL VISUAL RESULTS

We add more evalution result in the Figure 2 and Figure 3. For DM-based methods in Figure 3, we input the text prompt [Pyramid,

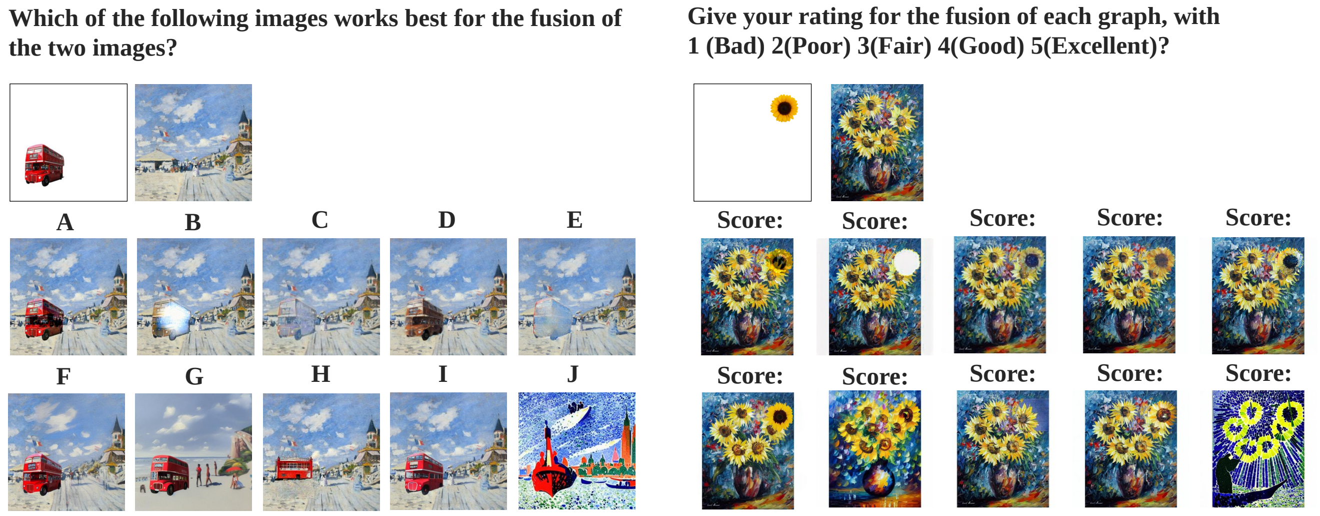
Which of the following images works best for the fusion of the two images?
A
B
C
D
E
F
G
H
I
J
Give your rating for the fusion of each graph, with 1 (Bad) 2(Poor) 3(Fair) 4(Good) 5(Excellent)?
Score:
Score:
Score:
Score:
Score:
Score:
Score:
Score:
Score:
Score:

Figure 1: *Left* **Questionnaire for top-1 vote.** *Right* **Questionnaire for score rating.**

Bus, Boat, Sphinx, Moai, Bench, Red fire hydrant, Television, Apple, Tree, Aircraft, Croissant] for SDEdit, SD-text, FreePIH from the first row to the last row. For SD-Text, we input the follow prompt to guide the diffusion to generate the images:

- A yellow boat is on the river. The sun is going down.
- A painting of a Sphinx in the countryside. The Sphinx is near a house and river.
- A moai statue in an oil painting with various colors, the moai statue is near a tree and a street light.
- A snow scene, a bench against a wall, a man walking in the middle.
- Three people are talking in the house. A TV is in the lower right corner.
- A table with a vase, five apples and a wine glass on it.
- A snow scene. There are a lot of trees on the left and a house on the right.
- A yellow plane is flying in the sky. There are a lot of people grazing on the ground, village view.
- An oil painting showing a fruit bowl with oranges and croissants.

REFERENCES

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended Diffusion for Text-driven Editing of Natural Images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 18187–18197. https://doi.org/10.1109/CVPR52688.2022.01767

[2] Junyan Cao, Yan Hong, and Li Niu. 2023. Painterly Image Harmonization in Dual Domains. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 268–276. https://ojs.aaai.org/index.php/AAAI/article/view/25099

[3] Roy Hachnochi, Mingrui Zhao, Nadav Orzech, Rinon Gal, Ali Mahdavi-Amiri, Daniel Cohen-Or, and Amit Haim Bermano. 2023. Cross-domain Compositing with Pretrained Diffusion Models. *arXiv preprint arXiv:2302.10167* (2023).

[4] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 8693)*, David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[5] Lingxiao Lu, Jiangtong Li, Junyan Cao, Li Niu, and Liqing Zhang. 2023. Painterly Image Harmonization using Diffusion Model. In *Proceedings of the 31st ACM International Conference on Multimedia*. 233–241.

[6] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=aBsCjcPu_tE

[7] Patrick Pérez, Michel Gangnet, and Andrew Blake. 2003. Poisson image editing. *ACM Trans. Graph.* 22, 3 (2003), 313–318. https://doi.org/10.1145/882262.882269

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 10674–10685. https://doi.org/10.1109/CVPR52688.2022.01042

[9] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*. http://papers.nips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html

[10] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. 2020. Deep Image Blending. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*. IEEE, 231–240. https://doi.org/10.1109/WACV45572.2020.9093632

[11] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10146–10156.

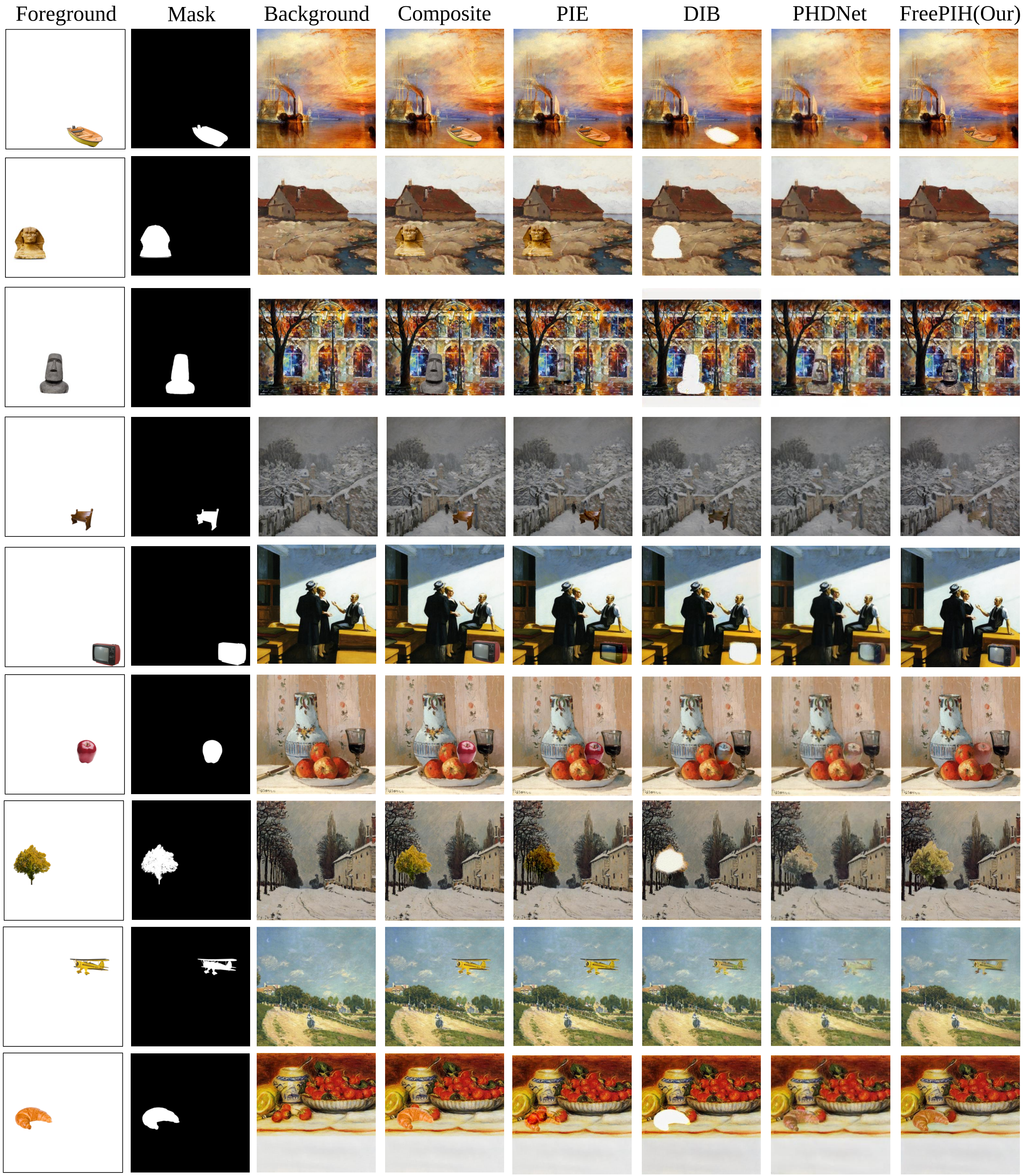

Figure 2: Evaluation results compared with non-DM based methods.

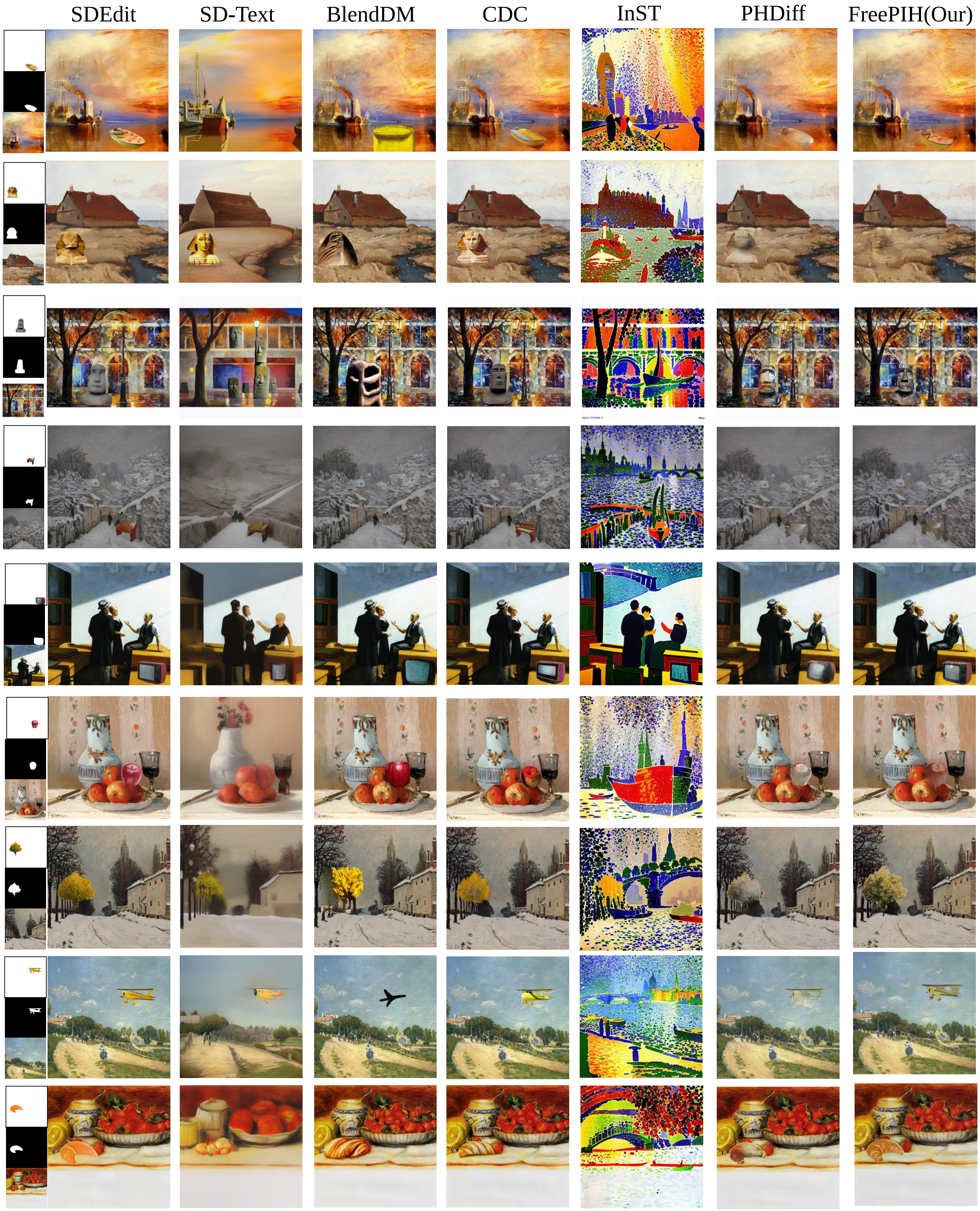

Figure 3: Evaluation results compared with DM based methods.