PTDiffusion: Free Lunch for Generating Optical Illusion Hidden Pictures with Phase-Transferred Diffusion Model

Supplementary Material

6. Preliminary background

6.1. Diffusion model background

Since the advent of Denoising Diffusion Probabilistic Model (DDPM), diffusion model has soon dominated research field of generative AI due to its advantages in training stability and sampling diversity as compared with GAN. Grounded in the theory of stochastic differential equations, diffusion model learns to iteratively denoise a noise-corrupted input signal (*e.g.*, an image or a video clip), ultimately generating clean data that follow the underlying target distribution. Diffusion model is conceptually composed of a forward diffusion process and a reverse denoising process. The forward diffusion process gradually adds noise to the data over a series of steps, transforming the data into a random Gaussian distribution, while the reverse denoising process learns to reverse the forward process by iteratively removing noise from the data, starting from pure noise and gradually reconstructing the original data. The model is trained to predict the noise added at each step of the forward process. By learning to denoise, the model can generate new data samples by starting from random noise and applying the reverse process.

Given the original data distribution $q(x_0)$, the forward diffusion process applies a T-step Markov chain to gradually add noise to the original data x_0 according to the conditional distribution $q(x_t|x_{t-1})$, which is defined as follows:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathcal{I}), \tag{15}$$

where α_t follows a predefined schedule, $\alpha_t \in (0,1)$, $\alpha_t > \alpha_{t+1}$. Using the notation $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, we can derive the marginal distribution $q(x_t|x_0)$ that can be used to directly obtain x_t from x_0 in a single step for arbitrary time step t:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathcal{I}), \tag{16}$$

where $\sqrt{\bar{\alpha}_T} \approx 0$. With the forward diffusion process, the source data x_0 is transformed into x_T that follows an isotropic Gaussian distribution.

The reverse denoising process learns to conversely convert a Gaussian noise x_T to the manifold of the original data distribution $q(x_0)$ by gradually estimating and sampling from the posterior distribution $p(x_{t-1}|x_t)$. Since the posterior distribution $p(x_{t-1}|x_t)$ is mathematically intractable, we can derive the conditional posterior distribution $p(x_{t-1}|x_t, x_0)$ with the Bayes formula and some algebraic manipulation:

$$p(x_{t-1}|x_t, x_0) = \mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t\mathcal{I}), \tag{17}$$

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t, \tag{18}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t, \tag{19}$$

where $\beta_t = 1 - \alpha_t$. However, the conditional posterior distribution $p(x_{t-1}|x_t, x_0)$ cannot be directly used for sampling since x_0 is unavailable at inference time (x_0 is the target of the sampling process). Thus, DDPM tries to estimate the unknown x_0 given the x_t at each time step. Considering the reparameterization form of Eq. 16:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \tag{20}$$

in which ϵ_t denotes the randomly sampled Gaussian noise that maps x_0 to x_t in a single step according to Eq. 16. Given Eq. 20, we can represent x_0 using x_t and ϵ_t:

$$x_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_t). \tag{21}$$

However, the Gaussian noise ϵ_t sampled in the forward diffusion process is also unknown for the reverse denoising process where only x_t is available. Consequently, DDPM builds a noise estimation network ϵ_θ that predicts the sampled Gaussian noise ϵ_t in Eq. 21 with x_t and time step t as input, which is realized by training ϵ_θ with the following noise regression loss:

$$L = \|\epsilon_t - \epsilon_\theta(x_t, t)\|_2, \tag{22}$$

where $t \sim \text{Uniform}(\{1, ..., T\})$, $\epsilon_t \sim \mathcal{N}(0, \mathcal{I})$, x_t is computed via Eq. 20. After model training, $y_\theta(x_t)$, the estimation of x_0 given x_t, can be obtained simply by replacing ϵ_t in Eq. 21 with the predicted noise $\epsilon_\theta(x_t, t)$:

$$y_\theta(x_t) = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)). \tag{23}$$

Replacing the unknown x_0 in Eq. 17 with its predicted estimation $y_\theta(x_t)$ given by Eq. 23, we can sample x_{t-1} based on x_t from the approximate posterior distribution $\mathcal{N}(x_{t-1}; \tilde{\mu}_t(x_t, y_\theta(x_t)), \tilde{\beta}_t\mathcal{I})$, and thus sample the ultimate x_0 step by step from the initial Gaussian noise x_T.

6.2. Conditional diffusion model

Taking the image generation task as an example, conditional diffusion model tackles conditional image synthesis by introducing additional condition c to the model to guide image generation (denoising) process. In this paradigm, the

condition signal c together with x_t and time step t are taken as input to the noise estimation network ϵ_θ, such that ϵ_θ is trained to conditionally predict the added Gaussian noise in the forward diffusion process, as supervised by the randomly sampled ϵ_t in Eq. 20. The training loss given by Eq. 22 is correspondingly updated as:

$$L = \|\epsilon_t - \epsilon_\theta(x_t, t, c)\|_2, \tag{24}$$

where $t \sim \text{Uniform}(\{1, ..., T\})$, $\epsilon_t \sim \mathcal{N}(0, \mathcal{I})$, x_t is computed via Eq. 20. After model training, the reverse sampling process is applied to generate new images from random Gaussian noise x_T, based on the step-by-step denoising according to the conditional posterior distribution given by Eq. 17, in which the unknown x_0 is approximated by the linear combination of x_t and the conditional noise estimation, *i.e.*, the $y_\theta(x_t)$ (the approximate x_0 estimated by x_t) in Eq. 23 is updated as:

$$y_\theta(x_t, c) = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(x_t, t, c)). \tag{25}$$

6.3. Denoising diffusion implicit model

Denoising diffusion implicit model (DDIM) is a variant of diffusion model that builds on the framework of DDPM but enables much more efficient sampling while maintaining high-quality generation results. DDIM can generate samples in significantly fewer steps compared with DDPM by modeling the reverse denoising process as a non-Markovian process and skipping the intermediate denoising steps.

DDIM is totally the same as DDPM in model training and only differs with DDPM in model inference, namely that DDIM can directly inherit the pre-trained DDPM model. To compute x_{t-1} from x_t in the reverse denoising (sampling) process, DDIM features a two-step deterministic denoising. In the first step, DDIM estimates an approximate x_0 based on x_t using Eq. 23. In the second step, DDIM computes x_{t-1} from the approximate x_0 using the forward diffusion in the form of Eq. 20:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}y_\theta(x_t) + \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon_{t-1}, \tag{26}$$

where $y_\theta(x_t)$ is given by Eq. 23. Considering that the ϵ_{t-1} in the above equation is the sampled Gaussian noise in the forward diffusion process, which is unknown in the reverse denoising process, we can replace ϵ_{t-1} with $\epsilon_\theta(x_{t-1}, t-1)$, the approximate ϵ_{t-1} estimated by the network ϵ_θ. Therefore, the Eq. 26 can be updated as:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}y_\theta(x_t) + \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon_\theta(x_{t-1}, t-1). \tag{27}$$

However, the $\epsilon_\theta(x_{t-1}, t-1)$ in the above equation is also unavailable since x_{t-1} is unknown (we only know x_t and want to compute x_{t-1}). Thus, we can further approximate $\epsilon_\theta(x_{t-1}, t-1)$ with $\epsilon_\theta(x_t, t)$ and arrive to the final DDIM sampling equation:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}y_\theta(x_t) + \sqrt{1 - \bar{\alpha}_{t-1}}\epsilon_\theta(x_t, t). \tag{28}$$

Eq. 28 shows that the reverse sampling process of DDIM is totally deterministic, namely, each starting Gaussian noise x_T yields a unique sampling result x_0.

Note that the above derived two-step sampling process of $x_t \to x_0 \to x_{t-1}$ also applies for $x_t \to x_0 \to x_{t+1}$. That is, a clean image x_0 can be deterministically inverted into a Gaussian noise through the following inversion process:

$$x_{t+1} = \sqrt{\bar{\alpha}_{t+1}}y_\theta(x_t) + \sqrt{1 - \bar{\alpha}_{t+1}}\epsilon_\theta(x_t, t). \tag{29}$$

The DDIM inversion given by Eq. 29 has wide applications in image editing and style transfer. For conditional image generation of DDIM, the $y_\theta(x_t)$ and $\epsilon_\theta(x_t, t)$ in Eq. 28 and Eq. 29 are updated to $y_\theta(x_t, c)$ and $\epsilon_\theta(x_t, t, c)$ respectively.

6.4. Latent diffusion model

Latent diffusion model (LDM) compresses images from high-dimensional pixel space into low-dimensional feature space via pre-trained autoencoder, and builds diffusion model based on the latent feature space, such that computational overhead for both training and inference can be dramatically lowered. The training of LDM is similar to Eq. 24 except that we use notation z to denote latent features:

$$L = \|\epsilon_t - \epsilon_\theta(z_t, t, c)\|_2, \tag{30}$$

where $\epsilon_t \sim \mathcal{N}(0, \mathcal{I})$, $z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_t$, $z_0 = E(x_0)$, E is the pre-trained image encoder. The reverse denoising process from $z_T \sim \mathcal{N}(0, \mathcal{I})$ to z_0 is the same as $x_T \sim \mathcal{N}(0, \mathcal{I})$ to x_0 in DDPM. After reverse denoising process, the denoised clean features z_0 is decoded by the pre-trained decoder D to yield the finally generated image x_0, *i.e.*, $x_0 = D(z_0)$. In LDM framework, the condition c could be the extracted image features that are concatenated with x_t as the input of ϵ_θ for image-to-image translation applications, and also could be the encoded textual features that are interacted with x_t with cross-attention layers inside ϵ_θ for text-to-image synthesis task.

7. More qualitative results

Below we showcase more qualitative results of our PTDiffusion as a supplement to the main text. In Fig. 13 and Fig. 14, we display more results of hidden content discernibility control realized by varying the async distance parameter d. In Fig. 15 and Fig. 16, we display more examples demonstrating the sampling diversity property of our method, namely generating diversified illusion pictures with fixed reference image and text prompt. Finally, we present more optical illusion hidden pictures generated by our method in Fig. 17 to Fig. 27.

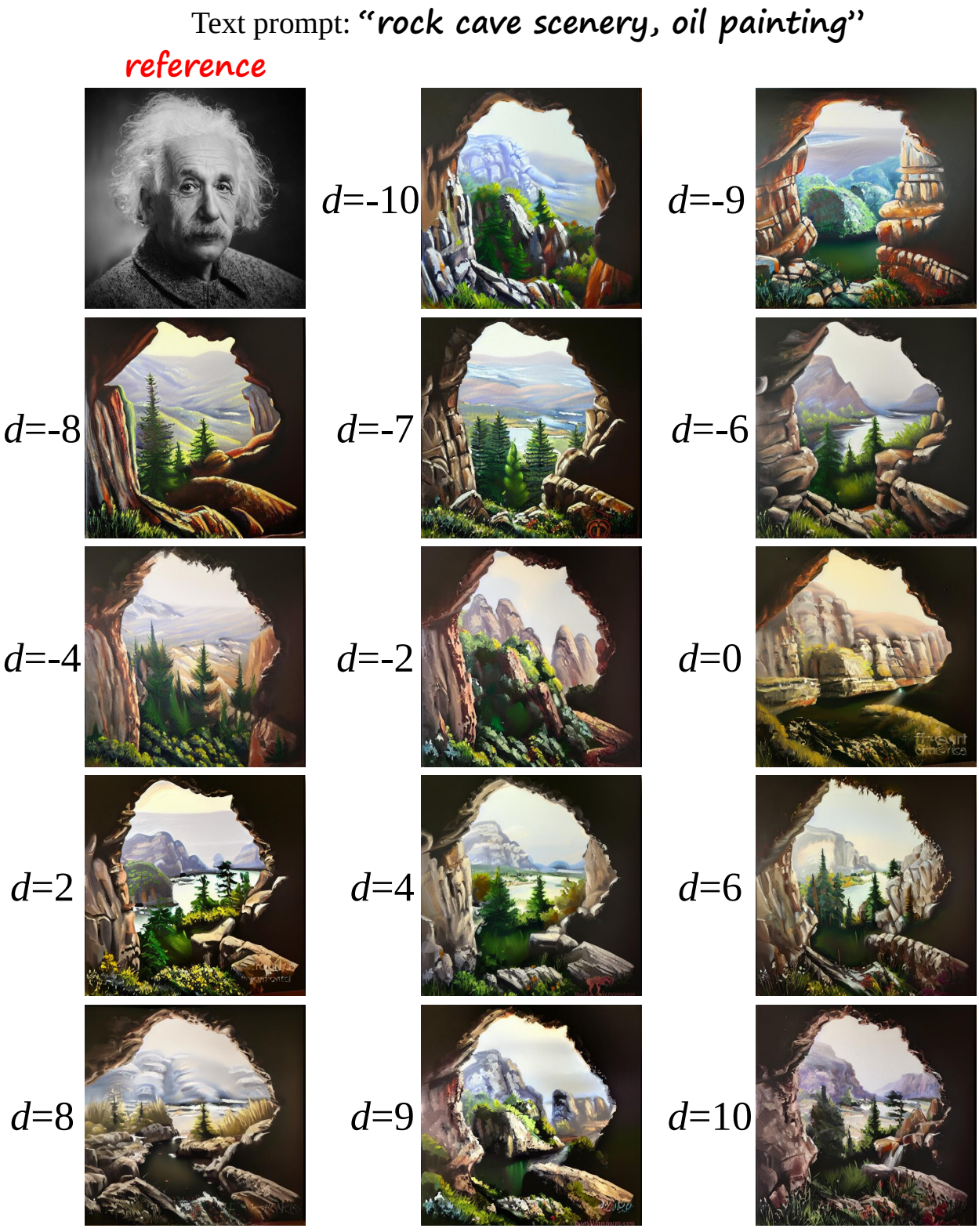

Figure 13. More results of hidden content discernibility control realized by varying the async distance parameter d in our method.

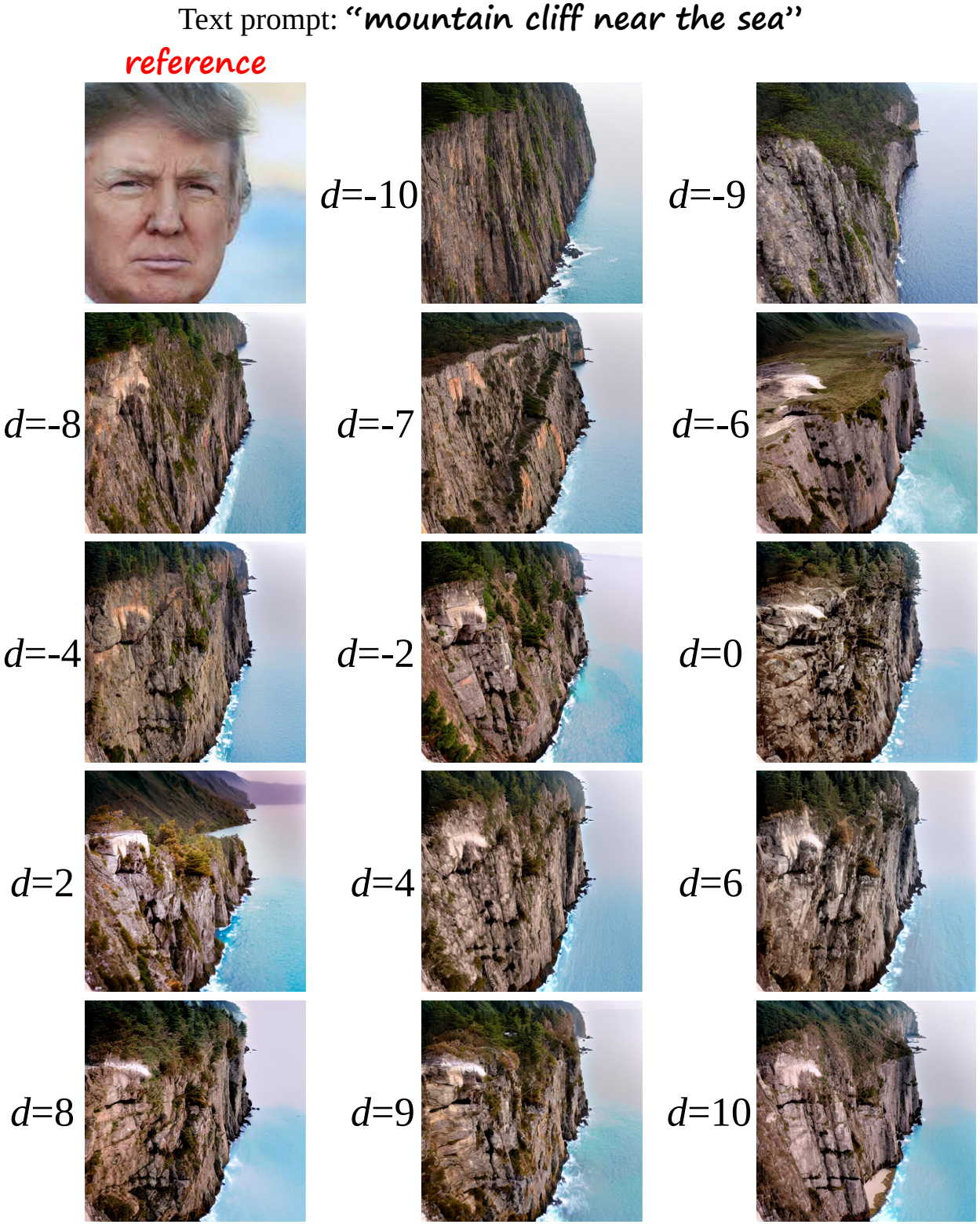

Figure 14. More results of hidden content discernibility control realized by varying the async distance parameter d in our method.

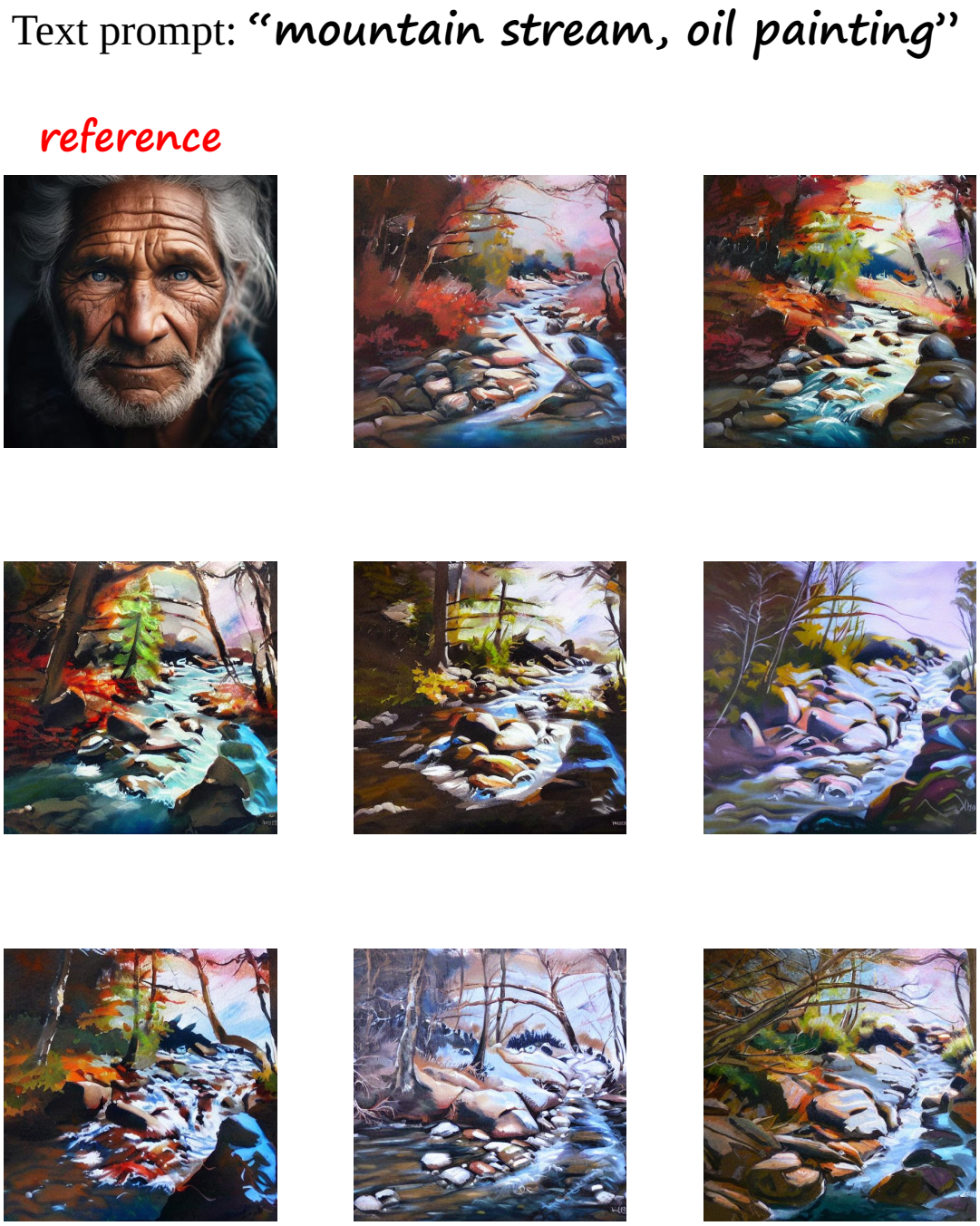

Figure 15. More examples of diversified sampling results of our method realized by varying the initial Gaussian noise $\tilde{z}_T$.

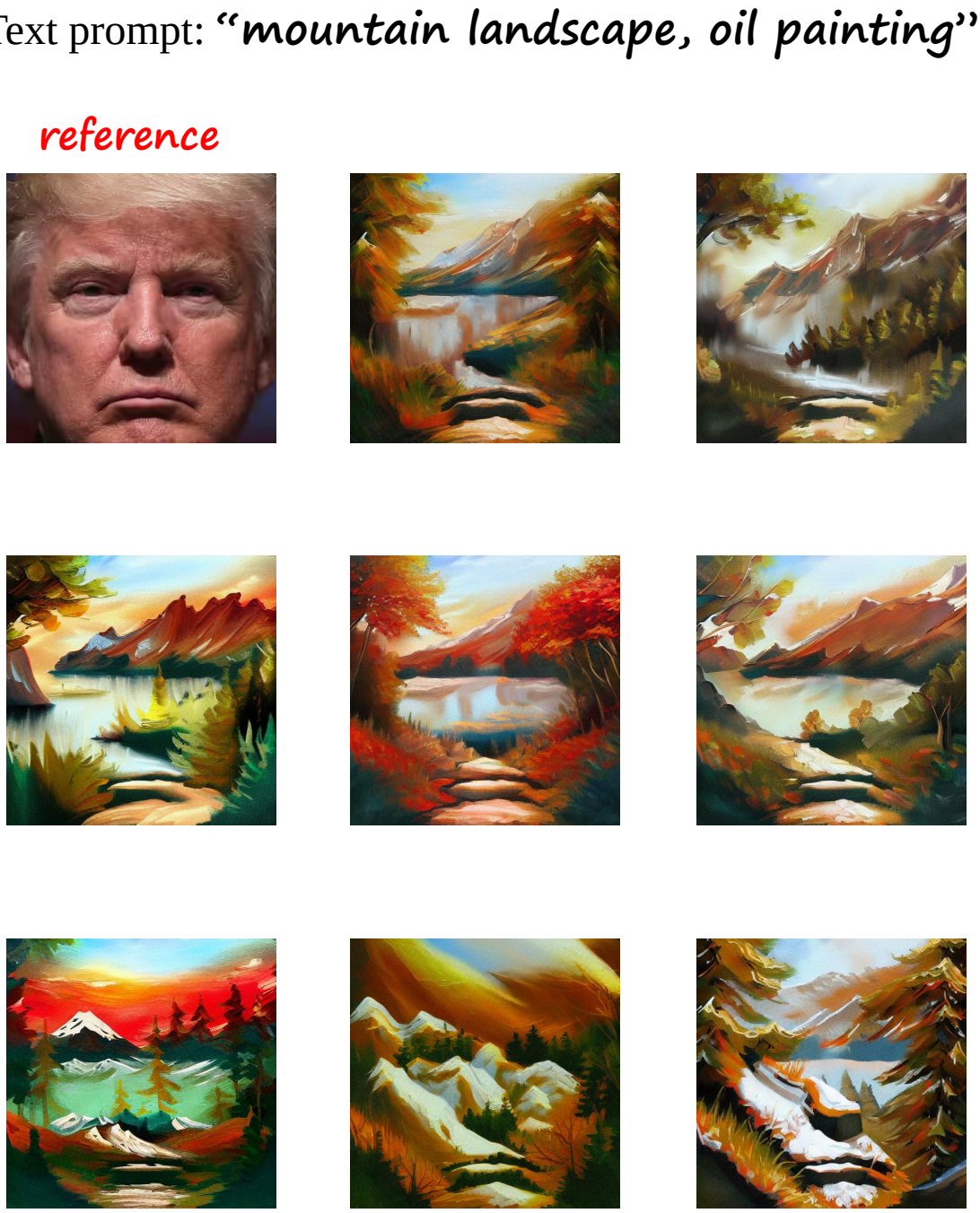

Figure 16. More examples of diversified sampling results of our method realized by varying the initial Gaussian noise $\tilde{z}_T$.

Figure 17. More qualitative results of our method.

Figure 18. More qualitative results of our method.

Figure 19. More qualitative results of our method.

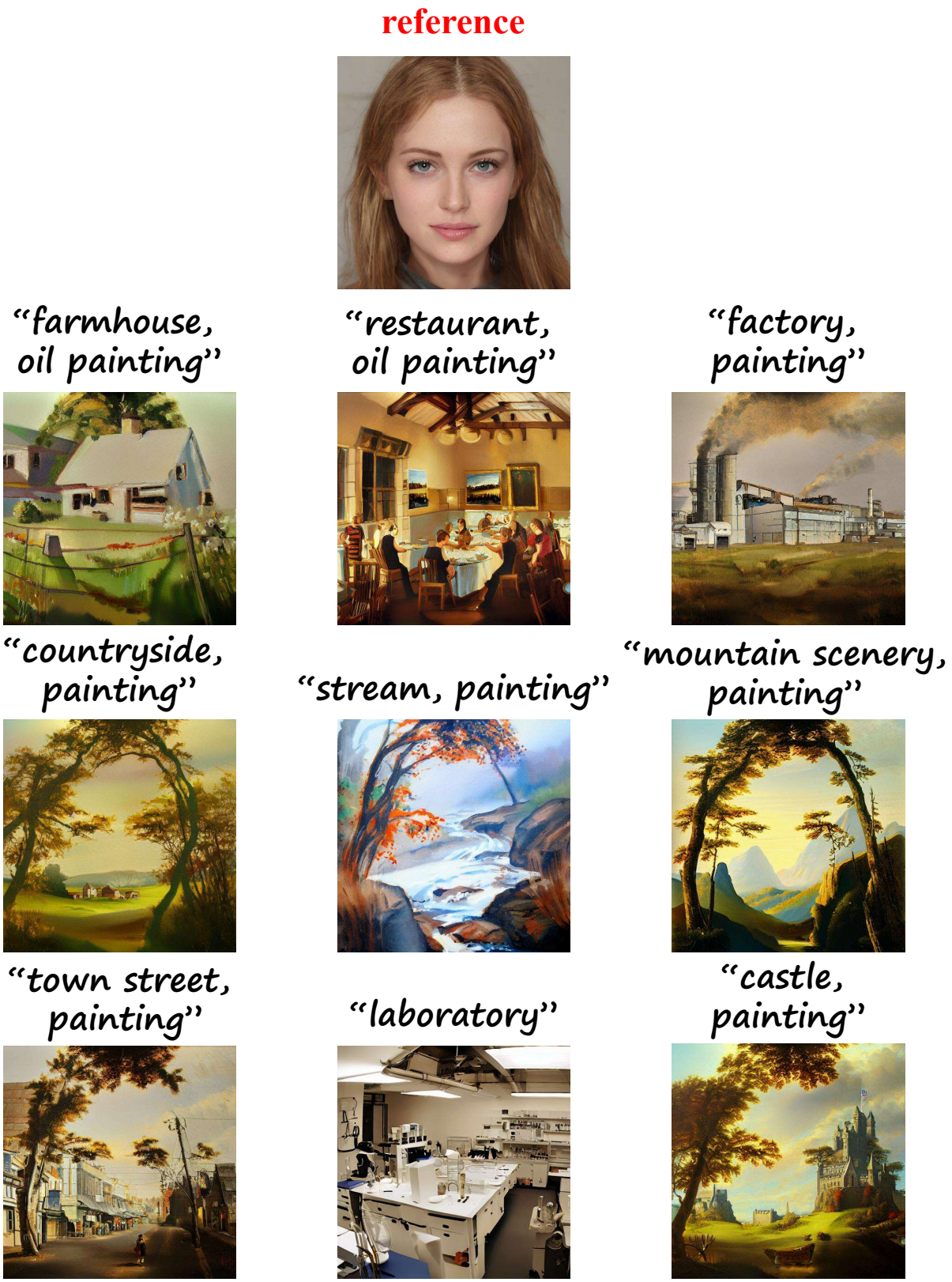

Figure 20. More qualitative results of our method.

Figure 21. More qualitative results of our method.

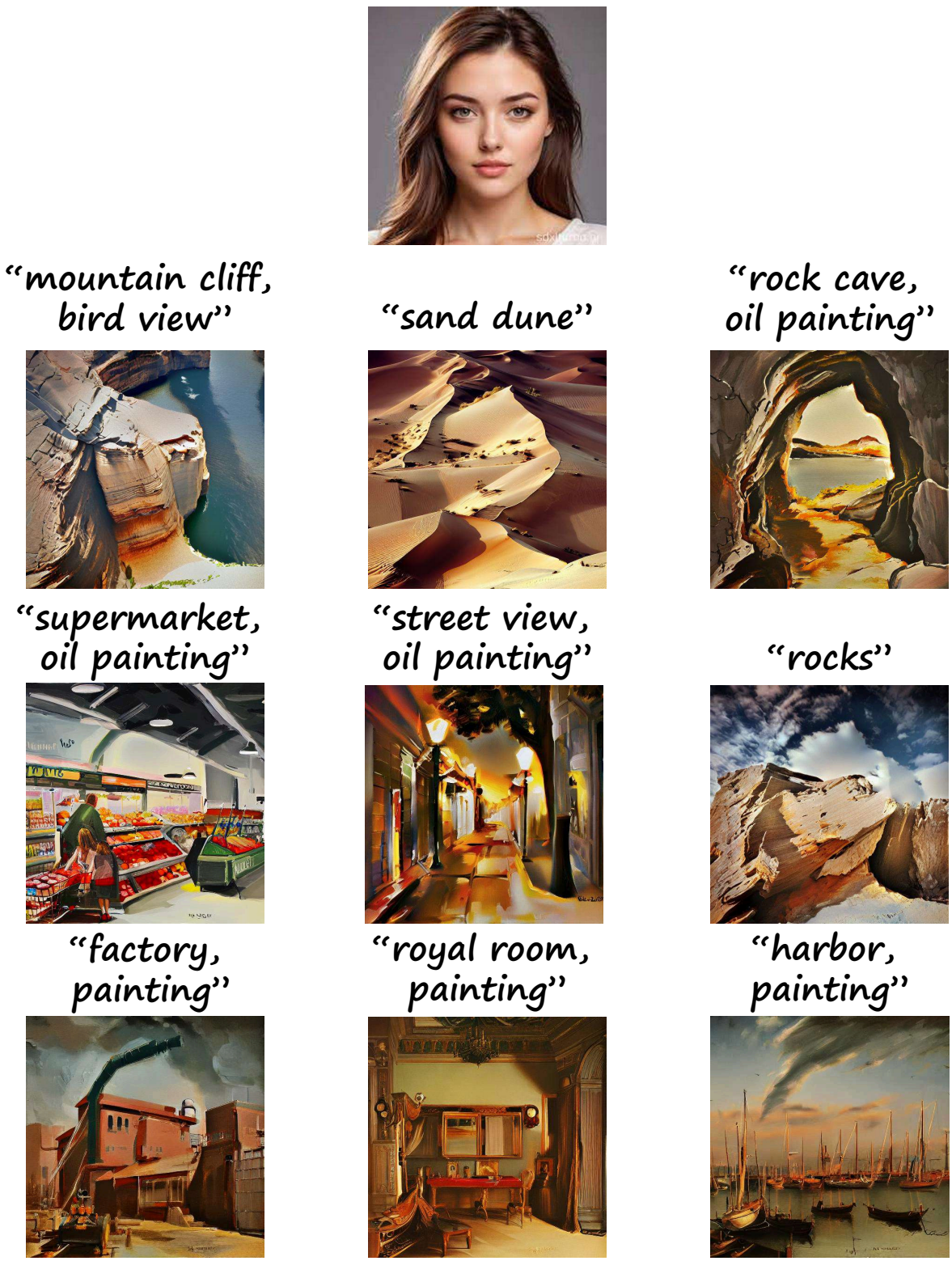

Figure 22. More qualitative results of our method.

Figure 23. More qualitative results of our method.

Figure 24. More qualitative results of our method.

Figure 25. More qualitative results of our method.

Figure 26. More qualitative results of our method.

Figure 27. More qualitative results of our method.

References

[1] Samaneh Azadi, Matthew Fisher, Vladimir G Kim, Zhaowen Wang, Eli Shechtman, and Trevor Darrell. Multi-content gan for few-shot font style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7564–7573, 2018. 3

[2] Ryan Burgert, Xiang Li, Abe Leite, Kanchana Ranasinghe, and Michael Ryoo. Diffusion illusions: Hiding images in plain sight. In *Proceedings of the ACM SIGGRAPH*, pages 1–11, 2024. 2

[3] Hung-Kuo Chu, Wei-Hsin Hsu, Niloy J Mitra, Daniel Cohen-Or, Tien-Tsin Wong, and Tong-Yee Lee. Camouflage images. In *Proceedings of the ACM SIGGRAPH*, pages 1–8. 2010. 2

[4] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 8780–8794, 2021. 2

[5] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7430–7440, 2023. 3, 5, 8

[6] Werner Ehm. A variational approach to geometric-optical illusions modeling. *Proceedings of Fechner Day*, 27(1):41–46, 2011. 2

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning*, 2024. 3

[8] William T Freeman, Edward H Adelson, and David J Heeger. Motion without movement. In *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, pages 27–30, 1991. 2

[9] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2427–2436, 2019. 3

[10] 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 ACM International Conference on Multimedia*, pages 4101–4109, 2024. 3, 6, 8

[11] Xiang Gao and Yuqi Zhang. Sragan: Saliency regularized and attended generative adversarial network for chinese ink-wash painting style transfer. *Pattern Recognition*, 162: 111344, 2025. 3

[12] Xiang Gao, Yingjie Tian, and Zhiquan Qi. Rpd-gan: Learning to draw realistic paintings with generative adversarial network. *IEEE Transactions on Image Processing*, 29:8706–8720, 2020. 3

[13] Xiang Gao, Yuqi Zhang, and Yingjie Tian. Learning to incorporate texture saliency adaptive attention to image cartoonization. In *International Conference on Machine Learning*, pages 7183–7207. PMLR, 2022. 3

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

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016. 2

[16] Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24154–24163, 2024. 2

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2, 3

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

[19] Elad Hirsch and Ayellet Tal. Color visual illusions: a statistics-based computational model. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 9447–9458, 2020. 2

[20] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 4

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 6840–6851, 2020. 2, 3

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3

[23] Yue Jiang, Zhouhui Lian, Yingmin Tang, and Jianguo Xiao. Scfont: Structure-guided chinese font generation via deep stacked networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4015–4022, 2019. 3

[24] Yuxin Jiang, Liming Jiang, Shuai Yang, and Chen Change Loy. Scenimefy: Learning to craft anime scene via semi-supervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7357–7367, 2023. 3

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 2

[26] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 3

[27] Dmytro Kotovenko, Artsiom Sanakoyeu, Pingchuan Ma, Sabine Lang, and Bjorn Ommer. A content transformation block for image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10032–10041, 2019. 3

[28] Hala Lamdouar, Weidi Xie, and Andrew Zisserman. The making and breaking of camouflage. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 832–842, 2023. 2

[29] Jia-Wei Liao, Winston Wang, Tzu-Sian Wang, Li-Xuan Peng, Ju-Hsuan Weng, Cheng-Fu Chou, and Jun-Cheng Chen. DiffQRCoder: Diffusion-based aesthetic qr code generation with scanning robustness guided iterative refinement. *arXiv preprint arXiv:2409.06355*, 2024. 2

[30] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 3, 6, 8

[31] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023. 2, 3, 5, 8

[32] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2, 3

[33] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 2, 3

[34] Aude Oliva, Antonio Torralba, and Philippe G Schyns. Hybrid images. *ACM Transactions on Graphics*, 25(3):527–532, 2006. 2

[35] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020. 3

[36] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *Proceedings of the ACM SIGGRAPH*, pages 1–11, 2023. 2, 3, 5

[37] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[38] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 3

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 3

[41] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *Proceedings of the ACM SIGGRAPH*, pages 1–10, 2022. 2

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Lit, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Raphael Gontijo-Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 36479–36494, 2022. 2, 3

[43] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[44] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 9

[45] 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*, pages 1921–1930, 2023. 2, 3, 5, 8

[46] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2D diffusion models for 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12619–12629, 2023. 2

[47] Guangyang Wu, Xiaohong Liu, Jun Jia, Xuehao Cui, and Guangtao Zhai. Text2QR: Harmonizing aesthetic customization and scanning robustness for text-guided QR code generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2024. 2

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3, 6, 8

[49] Qing Zhang, Gelin Yin, Yongwei Nie, and Wei-Shi Zheng. Deep camouflage images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12845–12852, 2020. 2

[50] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-ControlNet: all-in-one control to text-to-image diffusion models. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 11127–11150, 2023. 2

[51] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE*

international conference on computer vision, pages 2223–2232, 2017. 3