FreeDoM: Training-Free Energy-Guided Conditional Diffusion Model (Supplementary materials)

This supplementary document is organized as follows:

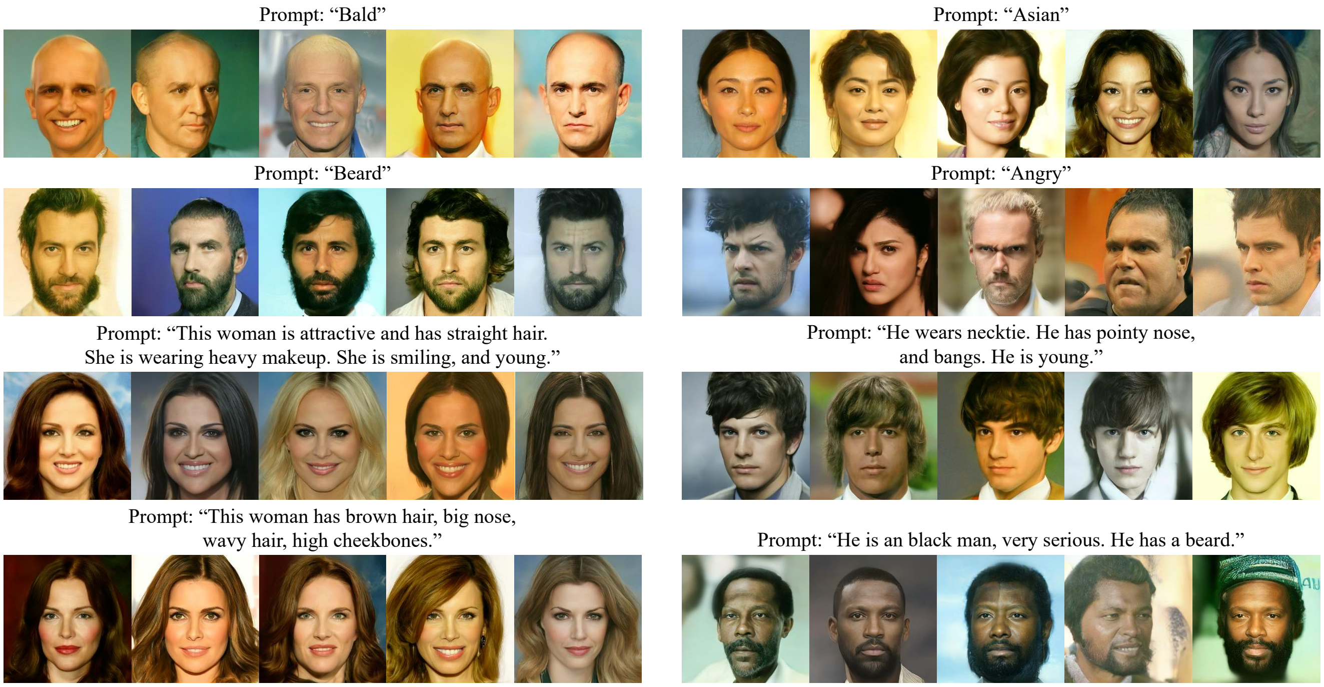
- Section 1: More results to show the performance of FreeDoM.
- Section 2: The detailed setting strategy of the learning rate ρ_t.
- Section 3: The setting details of the efficient time-travel strategy.
- Section 4: The relationship between FreeDoM and zero-shot image restoration methods.

1. More Results

In this section, we provide more generated results to demonstrate the effects of FreeDoM under various conditions and the applications FreeDoM support with training-required latent diffusion models.

We show the results of various conditions in Fig. 1 (text-to-image), Fig. 2 (segmentation-to-image), Fig. 3 (sketch-to-image), Fig. 4 (landmark-to-image), and Fig. 5 (id-to-image).

We show the results with latent diffusion models in Fig. 6 (style guidance + Stable Diffusion [7]), Fig. 7 (style guidance + Scribble ControlNet [12]) and Fig. 8 (face ID guidance + Human-pose ControlNet [12]). In order to further illustrate the implementation process of the application with the Human-pose ControlNet demonstrated in Fig. 8, we provide Fig. 9.



Prompt: “Bald”

Prompt: “Asian”

Prompt: “Beard”

Prompt: “Angry”

Prompt: “This woman is attractive and has straight hair. She is wearing heavy makeup. She is smiling, and young.”

Prompt: “He wears necktie. He has pointy nose, and bangs. He is young.”

Prompt: “This woman has brown hair, big nose, wavy hair, high cheekbones.”

Prompt: “He is an black man, very serious. He has a beard.”

Figure 1. Generated human faces for the text-to-image task. We choose four short and four long prompts to demonstrate the performance of FreeDoM. The characteristics described by these short prompts are experientially seldom seen in the training set. These results are consistent with the given conditions and have good diversity.

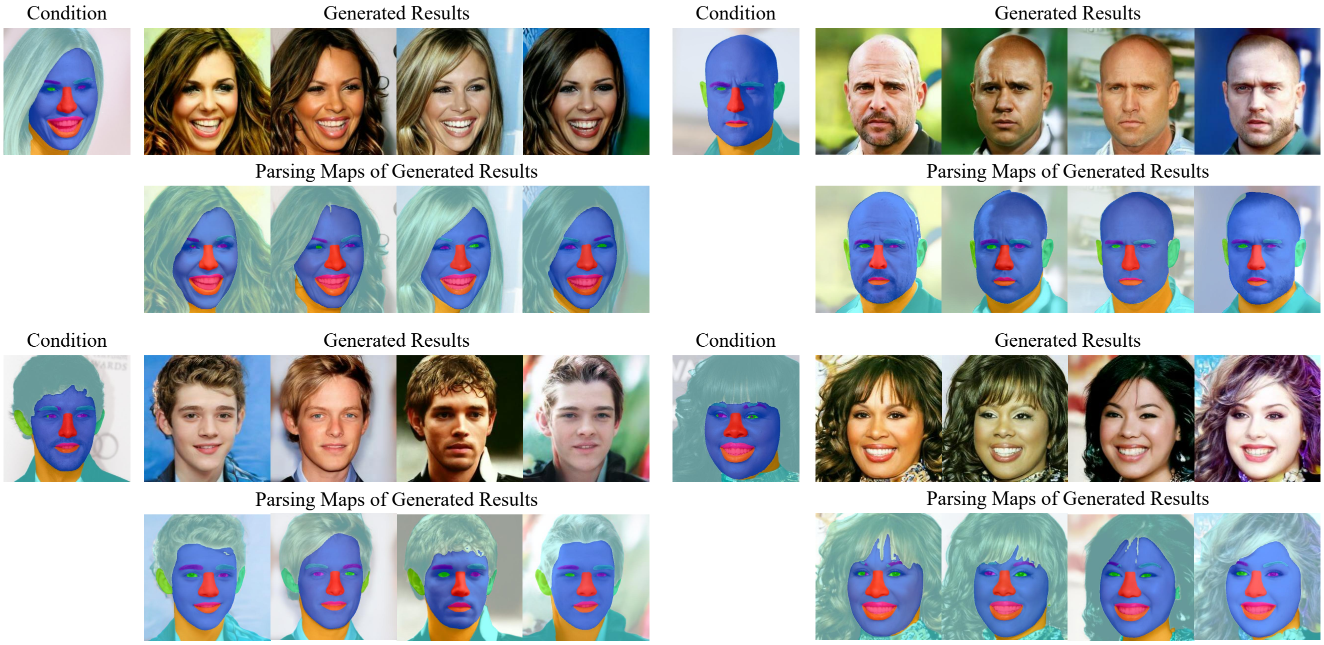

Condition
Generated Results
Condition
Generated Results
Parsing Maps of Generated Results
Parsing Maps of Generated Results
Condition
Generated Results
Condition
Generated Results
Parsing Maps of Generated Results
Parsing Maps of Generated Results

Figure 2. Generated human faces for the segmentation-to-image task. We choose four parsing maps to guide the generation process and output the parsing maps of the generated results to check the matching degree with given conditions. We can see that these results are consistent with the given conditions and have good diversity.

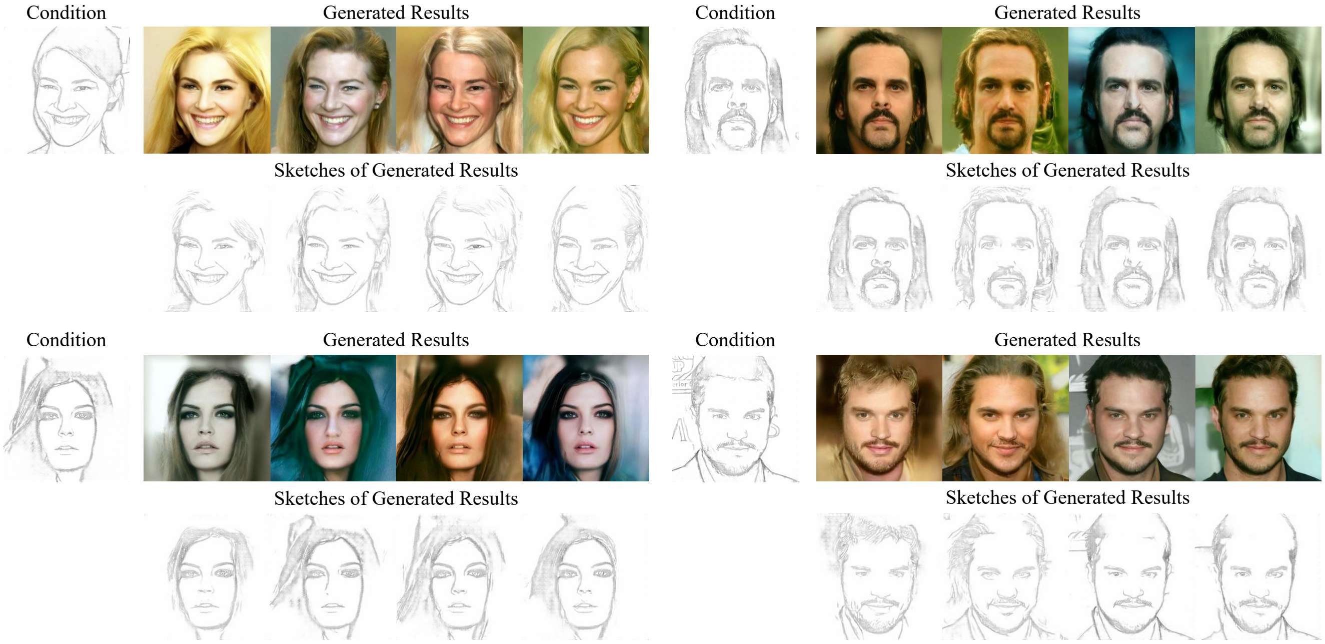

Condition
Generated Results
Condition
Generated Results
Sketches of Generated Results
Sketches of Generated Results
Condition
Generated Results
Condition
Generated Results
Sketches of Generated Results
Sketches of Generated Results

Figure 3. Generated human faces for the sketch-to-image task. We choose four sketches to guide the generation process and output the sketches of the generated results to check the matching degree with the given conditions. These results are consistent with the given conditions and have good diversity.

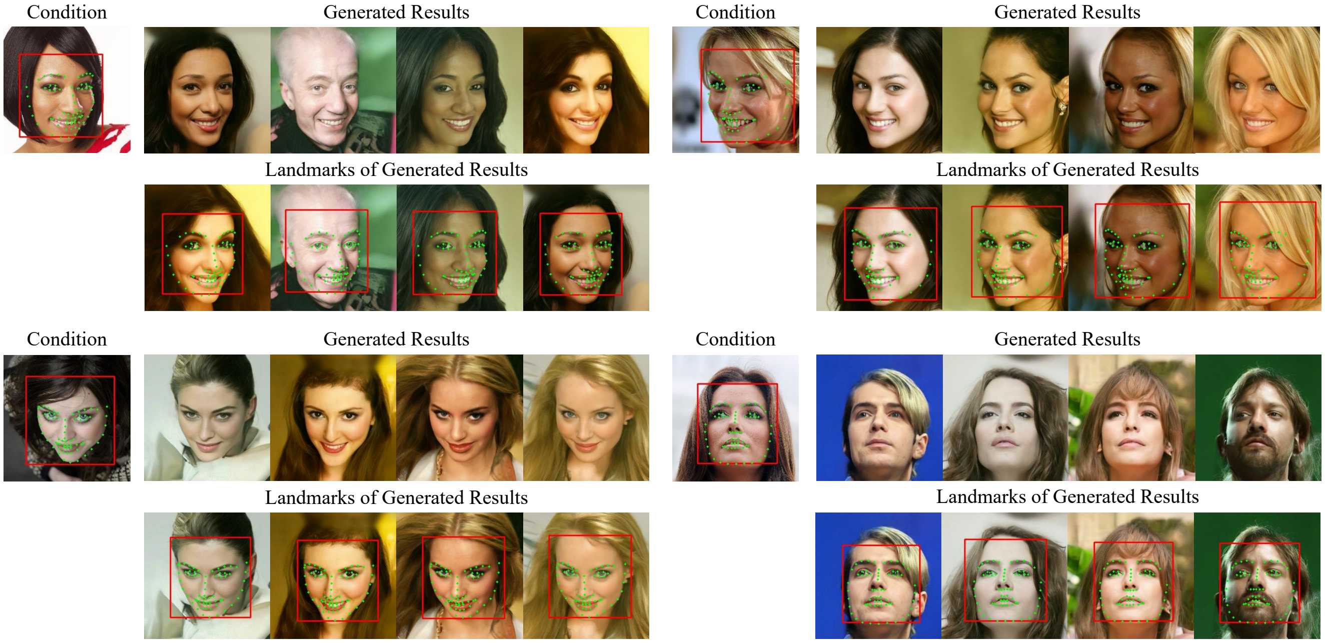

Condition
Generated Results
Condition
Generated Results
Landmarks of Generated Results
Landmarks of Generated Results
Condition
Generated Results
Condition
Generated Results
Landmarks of Generated Results
Landmarks of Generated Results

Figure 4. Generated human faces for the landmark-to-image task. We selected landmarks of four faces from different angles to guide the generation process and output the landmarks of the generated results to check the matching degree with given conditions. These results are consistent with the given conditions and have good diversity.

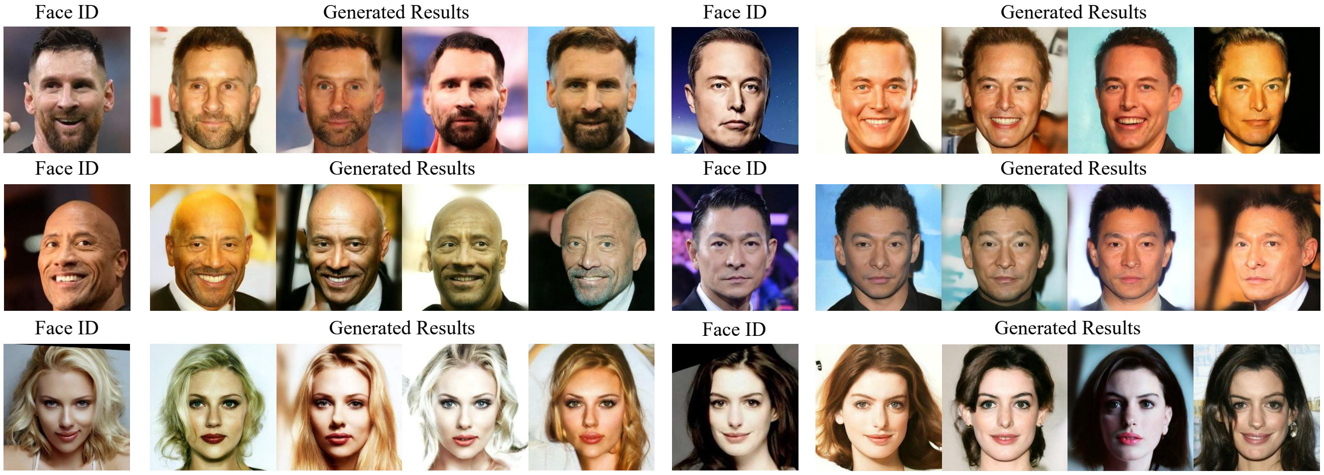


Figure 5. Generated human faces for ID-to-image task. We choose the face IDs of six celebrities as the reference to guide the generation process. These results are consistent with the given conditions and have good diversity.

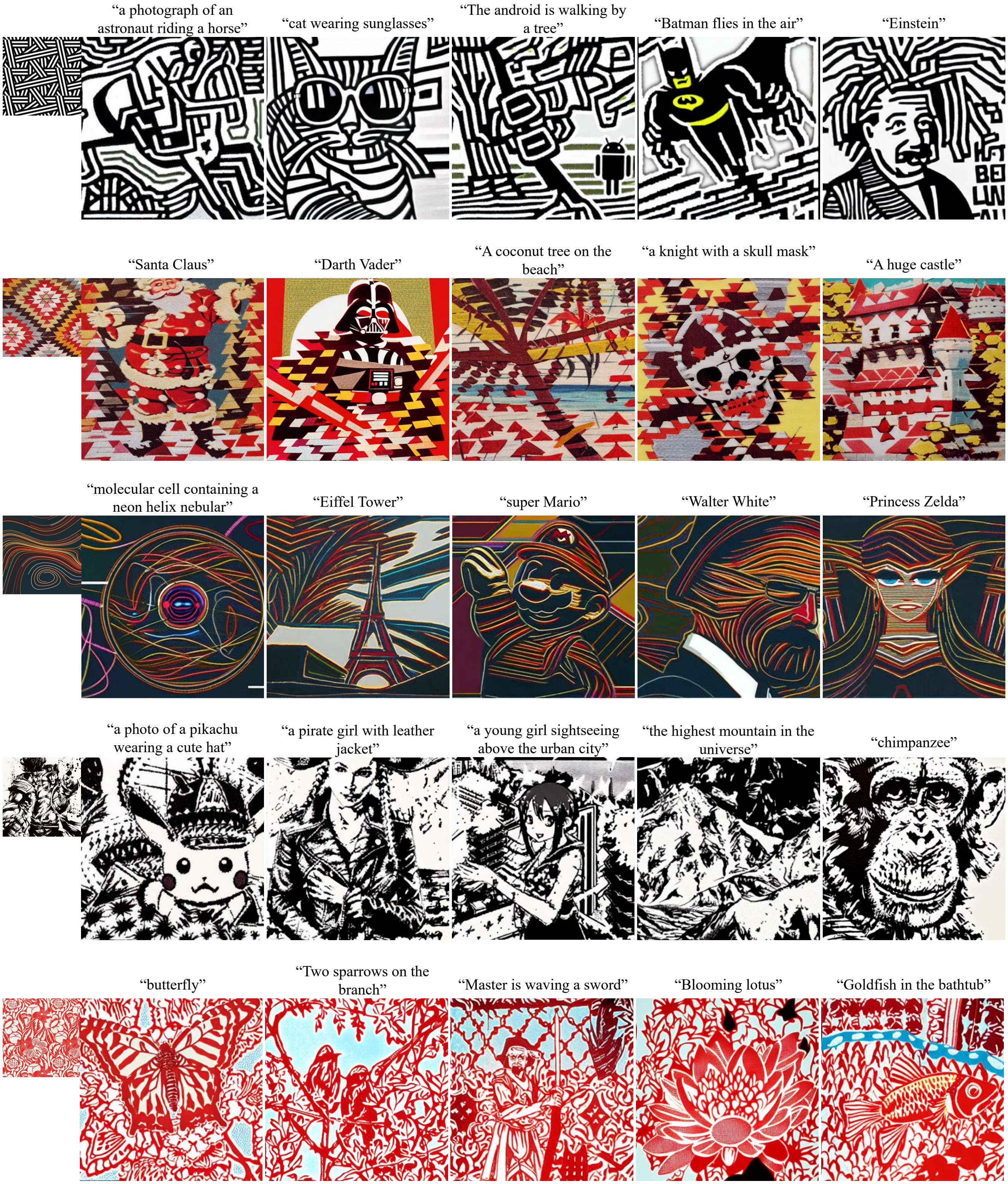

"a photograph of an astronaut riding a horse"
"cat wearing sunglasses"
"The android is walking by a tree"
"Batman flies in the air"
"Einstein"
"Santa Claus"
"Darth Vader"
"A coconut tree on the beach"
"a knight with a skull mask"
"A huge castle"
"molecular cell containing a neon helix nebular"
"Eiffel Tower"
"super Mario"
"Walter White"
"Princess Zelda"
"a photo of a pikachu wearing a cute hat"
"a pirate girl with leather jacket"
"a young girl sightseeing above the urban city"
"the highest mountain in the universe"
"chimpanzee"
"butterfly"
"Two sparrows on the branch"
"Master is waving a sword"
"Blooming lotus"
"Goldfish in the bathtub"

Figure 6. Generation results of training-free style guidance with text-to-image Stable Diffusion [7]. We choose five style images to guide the style of the results generated by Stable Diffusion. These generated results well match the provided style. **Zoom in for best view.**

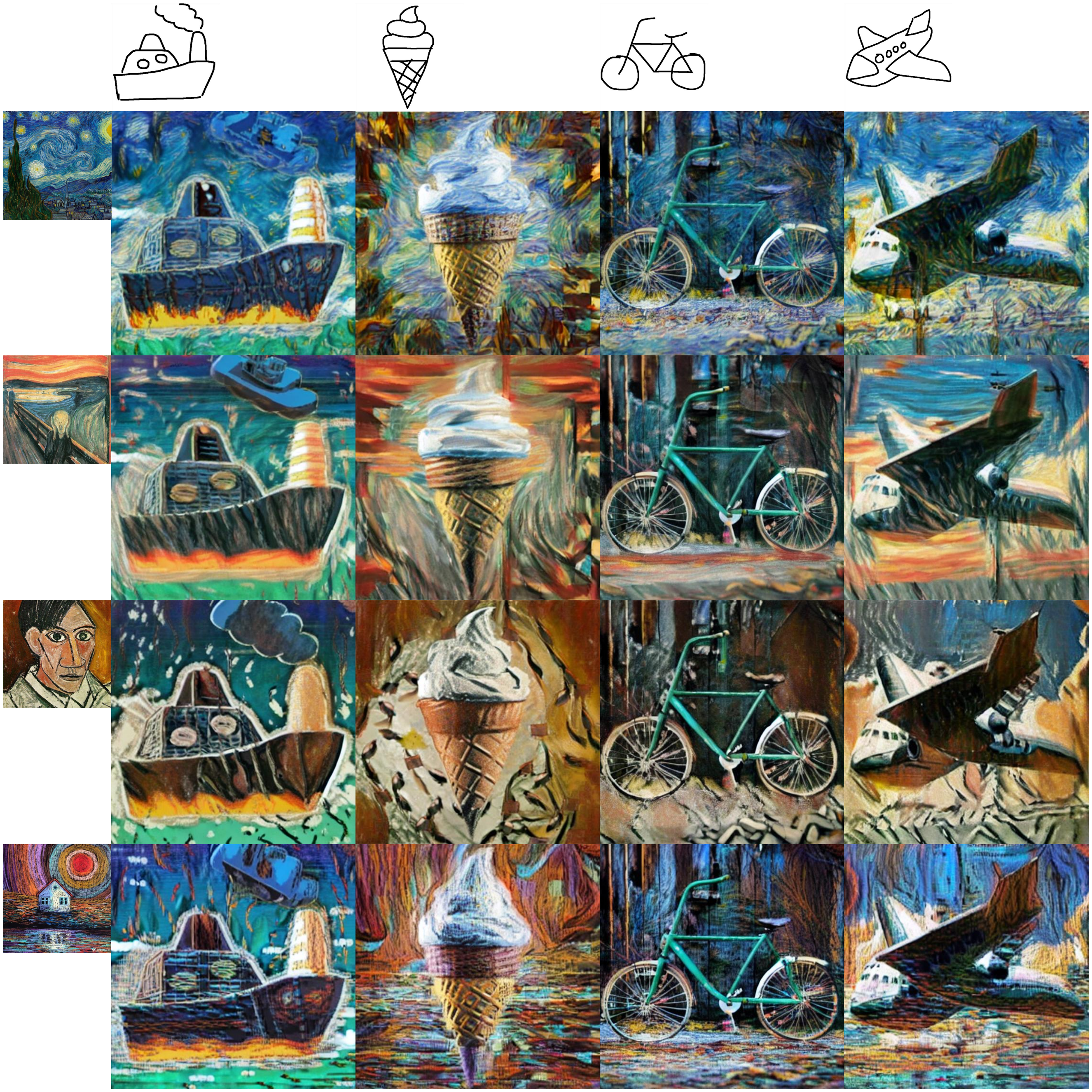


Figure 7. Generated results of training-free style guidance with Scribble ControlNet [12]. We choose four style images to guide the style of results generated by ControlNet. These generated results well match the provided style. **Zoom in for best view.**

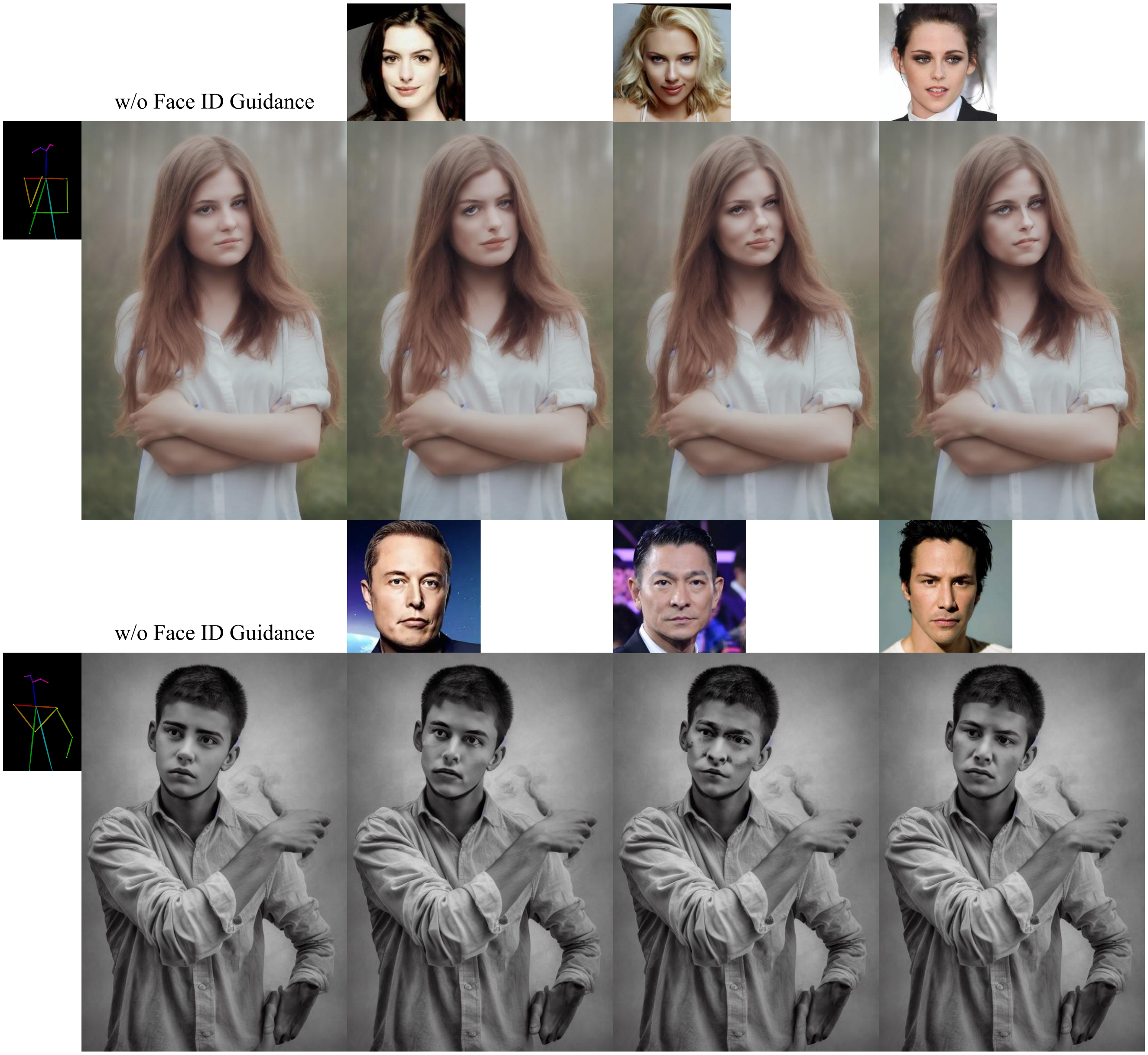


Figure 8. Generated Results of face ID guidance with Human-pose ControlNet [12]. By fixing random seeds, we can see the effects before and after introducing the ID guidance. These ID-guided results well match the given IDs in the face area. **Zoom in for best view.**

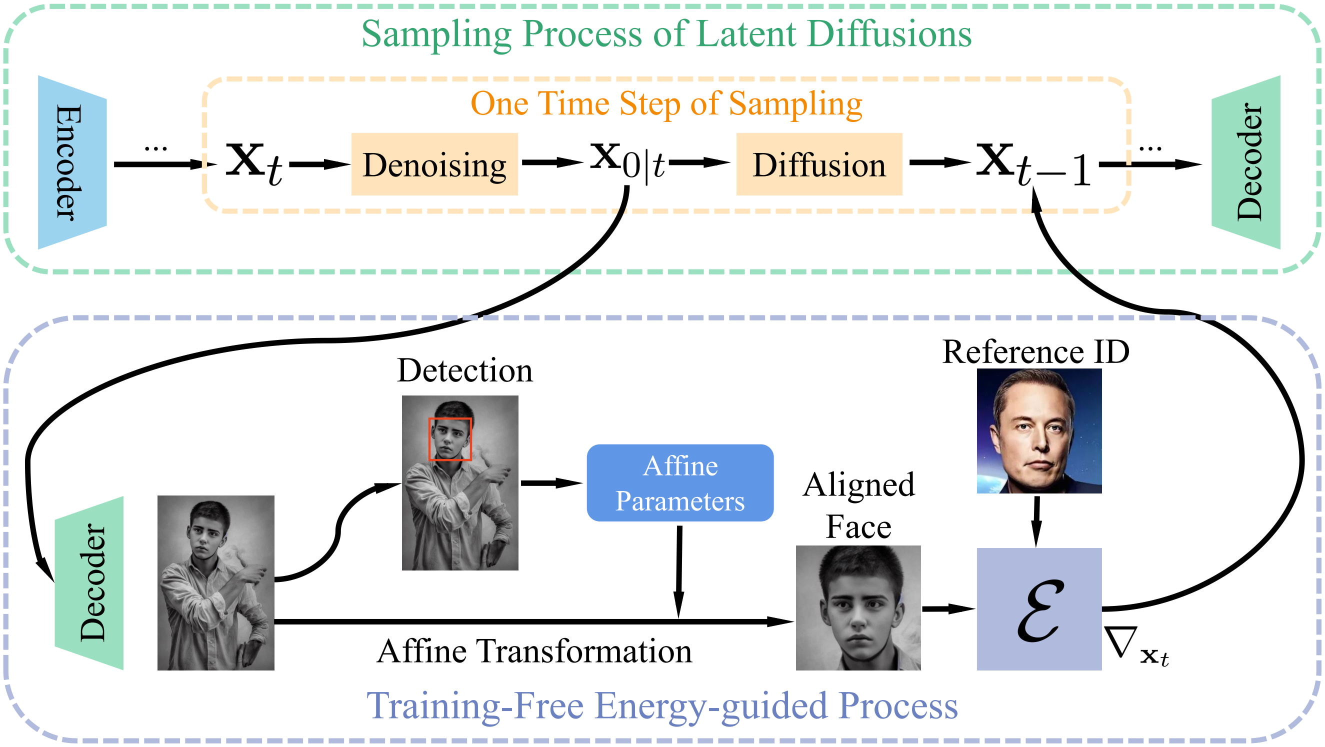


Figure 9. Visualization of the whole training-free face ID guidance process using FreeDoM in Fig. 8. We first decode the clean latent code $\mathbf{x}_{0|t}$ into the image domain. Then we detect the position of the human face and the corresponding landmarks. After getting the landmarks, we compute the affine parameters, which are used to perform an affine transformation to extract the aligned face area from the original decoded image. Finally, we compute the ID-based energy function between the aligned and reference faces. The gradient of the energy function to $\mathbf{x}_t$ will be used to update $\mathbf{x}_{t-1}$. Note that the computation of the Decoder and affine transformation is all differentiable, so the energy gradient to $\mathbf{x}_t$ is computable. **Zoom in for best view.**

2. Setting Strategy of Learning Rate

In the experiment, we found that the setting of the learning rate is the key to the effectiveness of FreeDoM. The best learning rate configuration and factors differ according to data domains, pre-trained models, and tasks. This section will introduce the relatively simple and effective learning rate setting methods for three situations summarized in our experiments for the community to reproduce.

- In the experiments on the ImageNet data domain, we try to determine the energy function gradient's learning rate according to the unconditional score's step size. This strategy ensures the stability of the generation process to avoid the collapse of results due to excessive step size. At the same time, this strategy only needs to adjust a factor that balances the unconditional score step size and the step size of the energy function gradient, which is convenient and feasible. The specific formula is as follows:

  $$\rho_t = 0.05 \frac{||\beta_t \cdot s(\mathbf{x}_t, t)||_2}{||\boldsymbol{g}_t||_2}, \tag{1}$$

  where $s(\mathbf{x}_t, t)$ predicts the unconditional score, β_t is the pre-defined parameter and $\boldsymbol{g}_t$ denotes the energy function gradient.

- For the experiments of latent diffusion models (Stable Diffusion [7] and ControlNet [12]), since they are classifier-free models, a better strategy is to determine the learning rate of the energy function gradient by referring to its conditional text-guided step size. In classifier-free methods, the conditional score function is computed as:

  $$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{c}) \approx s(\mathbf{x}_t, t, \emptyset) + r \cdot (s(\mathbf{x}_t, t, \mathbf{c}) - s(\mathbf{x}_t, t, \emptyset)), \tag{2}$$

  where the score predictor has a conditional input $\mathbf{c}$ and allows this input to be null (denoted as $\emptyset$). The factor r denotes the learning scale of the conditional guidance. In order to determine the learning rate in this situation, the specific formula is as follows:

  $$\rho_t = 0.2 \frac{||r \cdot \beta_t \cdot (s(\mathbf{x}_t, t, \mathbf{c}) - s(\mathbf{x}_t, t, \emptyset))||_2}{||\boldsymbol{g}_t||_2}, \tag{3}$$

  where β_t is the pre-defined parameter and $\boldsymbol{g}_t$ denotes the energy function gradient.

- For the human face experiments, we choose a simple but effective learning rate setting strategy, $\rho_t = k \cdot \sqrt{\bar{\alpha}_t}$ and k is different for each type of conditions. The specific values of k are shown in Tab. 1.

conditions	texts	segmentation maps	sketches	landmarks	face IDs
k	100	0.2	20	500	100

Table 1. The value of k under different conditions, which is an experimental choice.

In order to avoid artifacts in the final results, we will stop the guidance in the late part of the refinement stage (around 200-th to 1-th time step), which is a helpful trick to get satisfactory results.

3. Setting Details of the Efficient Time-Travel Strategy

In this section, we will introduce the configuration details related to the efficient time-travel strategy:

- In all the experiments of human faces, we found that the algorithm without the time-travel strategy can get enough satisfactory results, so we do not use the time-travel strategy.
- In the experiment on the ImageNet data domain, we only use the efficient time-travel strategy with $r_t = 10$ during the semantic stage between 800-th and 500-th time steps.
- In the experiment based on Stable Diffusion [7] and ControlNet [12] with style guidance, We only use the efficient time-travel strategy with $r_t = 3$ during the semantic stage between 800-th and 500-th time steps.
- An interesting discovery is that in the ControlNet-based face ID guidance experiments, we only need to add guidance without the efficient time-travel strategy in the refinement stage (between 500-th and 1-th time steps) to get acceptable results. A reasonable explanation is that the modification of face ID belongs to the modification of detail texture, which is the responsibility of the refinement stage. By dividing the whole sampling process into different stages, we can not only have a deeper understanding of the functions of each stage but also use this understanding to accelerate the overall sampling.

4. Relationship between FreeDoM and Zero-Shot Image Restoration Methods

The proposed FreeDoM is a framework that can support various conditions, including the degraded images in the image restoration tasks. Many existing zero-shot image restoration methods [1, 2, 3, 4, 5, 6, 8, 9, 11, 10] can be considered special cases of FreeDoM. Their idea can be summarized as updating the clean intermediate result $\mathbf{x}_{0|t}$ to meet the data consistency constraint, $\mathbf{y} = \mathcal{A}(\mathbf{x}_{0|t})$, where $\mathbf{y}$ is a degraded image and $\mathcal{A}(\cdot)$ is a linear or non-linear degradation operator. When dealing with linear degradation, the degradation operator $\mathcal{A}(\cdot)$ can be written into a matrix $\mathbf{A}$.

Since the image restoration tasks can also be seen as particular conditional generation tasks, these zero-shot image restoration methods can also be explained using the framework of FreeDoM. Take two typical examples: DPS [2] uses $-\nabla_{\mathbf{x}_t} ||\mathbf{y} - \mathcal{A}(\mathbf{x}_{0|t})||_2^2$ to update the intermediate results, which can be interpreted as a distance measurement function without learning parameters to improve the matching degree between the restored image $\mathbf{x}_{0|t}$ and the degraded image $\mathbf{y}$ in the measurement space; DDNM [11] obtains that the update direction for linear noiseless tasks is $-\mathbf{A}^{\dagger}(\mathbf{A}\mathbf{x}_{0|t} - \mathbf{y})$ through the derivation of Range-Null Space Decomposition, which can also be interpreted as an approximated analytical solution of the gradient of the distance measurement function in DPS on linear cases.

References

[1] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 9

[2] Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. 9

[3] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 9

[4] Hyungjin Chung, Byeongsu Sim, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 9

[5] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 9

[6] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 9

[7] 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 (CVPR)*, 2022. 1, 4, 8

[8] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023. 9

[9] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations (ICLR)*, 2021. 9

[10] Yinhuai Wang, Jiwen Yu, Runyi Yu, and Jian Zhang. Unlimited-size diffusion restoration. *arXiv preprint arXiv:2303.00354*, 2023. 9

[11] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations*, 2023. 9

[12] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023. 1, 5, 6, 8