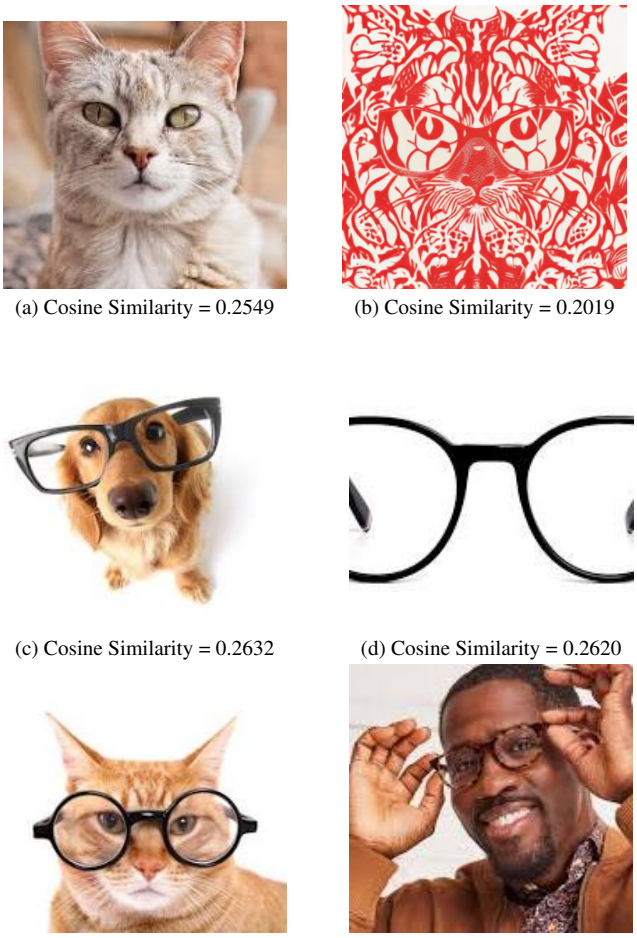
Z-Magic: Zero-shot Multiple Attributes Guided Image Creator (Supplementary Materials)

Yingying Deng, Xiangyu He[1],Fan Tang[✉,2], Weiming Dong[1]
[1] MAIS, Institute of Automation, Chinese Academy of Sciences
[2] Institute of Computing Technology, Chinese Academy of Sciences
dyy15@outlook.com, tfan.108@gmail.com, weiming.dong@ia.ac.cn

1. Discussions on Limitations

As a conditional generation method built on a score-based diffusion model, our approach also inherits the limitation of a higher sampling time cost. While we alleviate the dependency on hand-crafted training pair designs, the inference process remains slower compared to training-based methods. This slowdown arises primarily from the need to compute an additional derivative for the energy function during each iteration, which increases the overall computational cost. Previous works have demonstrated that strategies like time-travel [7] or corrector sampling [6] can enhance performance. To ensure a fair comparison, we adopt the same settings in our approach. However, adopting these strategies necessitates additional sampling steps, further contributing to the computational overhead. Although these enhancements align with best practices and ensure a fair comparison with other approaches, they amplify the time cost, highlighting the trade-off between improved performance and computational efficiency in our method.

Another significant challenge lies in the selection of the condition classifier $p(\boldsymbol{c}|\boldsymbol{x}_t)$. To meet the requirements of a zero-shot setting, we employ training-free classifiers, which leverage pre-trained discriminative models designed for tasks such as segmentation, landmark detection, and text-image coherence. These models enable flexibility and eliminate the need for task-specific fine-tuning. However, during our experiments, we observed that the loss values computed by these classifiers can sometimes diverge from human visual perception, posing a limitation in achieving optimal results. This inconsistency is particularly evident in widely used loss functions such as ArcFace loss [1] for identity preservation and CLIP loss [4] for object creation. For instance, CLIP loss, while effective for measuring general text-image alignment, struggles to capture nuanced differences in stylized or non-photorealistic images. As demonstrated in Figure 1, the cosine similarity scores produced by CLIP show a diminishing ability to reflect actual perceptual quality when applied to highly stylized or



(a) Cosine Similarity = 0.2549 (b) Cosine Similarity = 0.2019

(c) Cosine Similarity = 0.2632 (d) Cosine Similarity = 0.2620

(e) Cosine Similarity = 0.3179 (f) Cosine Similarity = 0.1925

Figure 1. We visualize the cosine similarity scores calculated by CLIP for the given images using the prompt "cat wearing glasses." The results reveal that CLIP performs poorly on stylized images, highlighting a significant gap between real-world images and the generated conditional outputs.

✉Corresponding author: Fan Tang.

Figure 2. We visualize the results of our approach conditioned on a pose prior and two different face IDs, thereby constructing a creation process guided by three attributes. While the generated male actor appears visually more similar to the given face photo compared to the female actor, the ArcFace loss is 0.44, which is higher than the female actor's loss of 0.37. This inconsistency hidden in ArcFace loss limits our approach in generating high-fidelity images.

abstract images, underscoring a notable gap between computational metrics and human judgment. Similarly, ArcFace loss may saturate and fail to distinguish subtle identity variations once the similarity score exceeds a certain threshold, as shown in Figure 2. While our method excels at optimizing multi-attribute outputs using these loss functions, its performance is inherently constrained by the limitations of the underlying metrics. The development or integration of more perceptually aligned loss functions, particularly those tailored for specific conditions or stylistic domains, could significantly enhance the quality and consistency of the generated results.

Finally, as our method heavily depends on various pre-trained open-source diffusion models for conditional generation, the fidelity and quality of the generated outputs are intrinsically linked to the performance of these underlying base models. This dependency implies that the capabilities of our approach are inherently constrained by the strengths and limitations of the pre-trained models we employ. As a result, the Fréchet Inception Distance (FID) scores [2] for our method may appear relatively lower when compared to state-of-the-art, domain-specific models that are fine-tuned for specific target tasks. Fine-tuned models are optimized for particular datasets or objectives, often enabling them to achieve superior fidelity and task-specific performance. Moreover, our method may not be ideal for tasks that cannot be effectively represented or optimized using a well-defined loss function. Without a robust and task-aligned loss metric, it becomes challenging to guide the generation process toward the desired outcomes, potentially limiting the applicability of our approach in certain scenarios.

Figure 3. We visualize the stylized results guided by prompts incorporating both style and text conditions. Eight images are selected, with all generated images guided by the same style reference image.

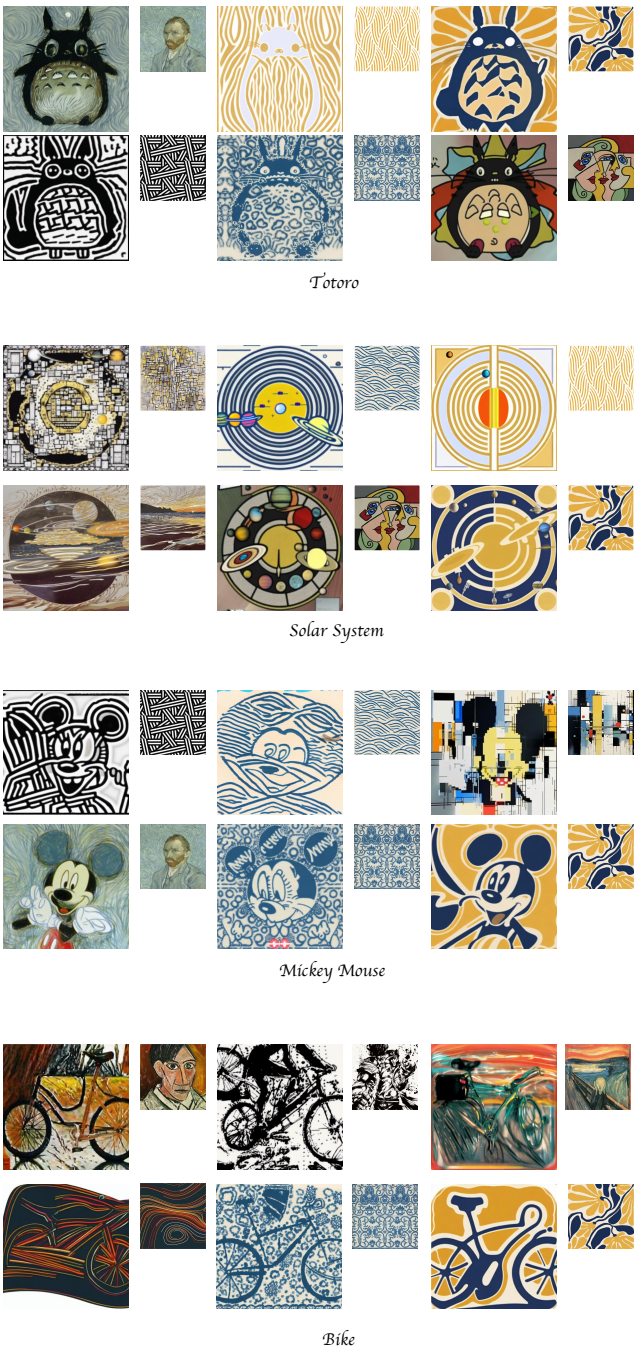


Figure 4. We visualize the stylized results guided by prompts using style and text conditions. Six images are selected, all generated from the same prompt, that closely adhere to the style reference images.

2. More Visualization Results

In this section, we present additional visualization results to illustrate the effectiveness of our approach when applied to various attributes.

2.1. Style Transfer

We incorporate the reference style $\mathbf{c}_1$ into the prompt-guided images generated by the Stable Diffusion v1.4 model [5], which serves as $p(\mathbf{x}|\mathbf{c}_0)$. The results, presented in Figures 3 and 4, demonstrate that our approach effectively captures highly abstract artistic styles, accurately matching the patterns, color schemes, and brushstrokes characteristic of the artist. Examples include a bicycle rendered in the style of Vincent Willem van Gogh and Mickey Mouse depicted in the style of Piet Cornelies Mondrian.

2.2. Face Synthesis

We provide additional results for the face synthesis task, incorporating multiple attributes based on the open-source face diffusion model [3] and the pose-guided ControlNet framework [8]. Figure 5 showcases the synthesis results under two conditions: face identity and landmarks. These examples highlight the effectiveness of our method in preserving facial identity while adhering to specified structural constraints. Similarly, Figure 6 demonstrates the results when face segmentation maps and textual prompts are used to guide the synthesis process. This setup emphasizes the ability to integrate segmentation-based structural information with semantic control via text prompts.

For the three-condition control scenario, Figure 8 presents extensive examples illustrating the generation of new results based on the combined input of face identity, face parsing maps, and prompt guidance. These results further validate the robustness and versatility of our approach in handling complex attribute combinations for conditional face synthesis. We also present generated group photos conditioned on pose and two distinct face identities, involving three conditions in total in Figure 7. By utilizing different face IDs, our approach demonstrates its capability to effectively preserve the pose priors while accurately reconstructing the target faces. This showcases the model's ability to handle complex multi-condition scenarios, ensuring both structural consistency and identity fidelity in the generated group images.

References

[1] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):5962–5979, 2022. 1

[2] 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 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. 2

[3] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun

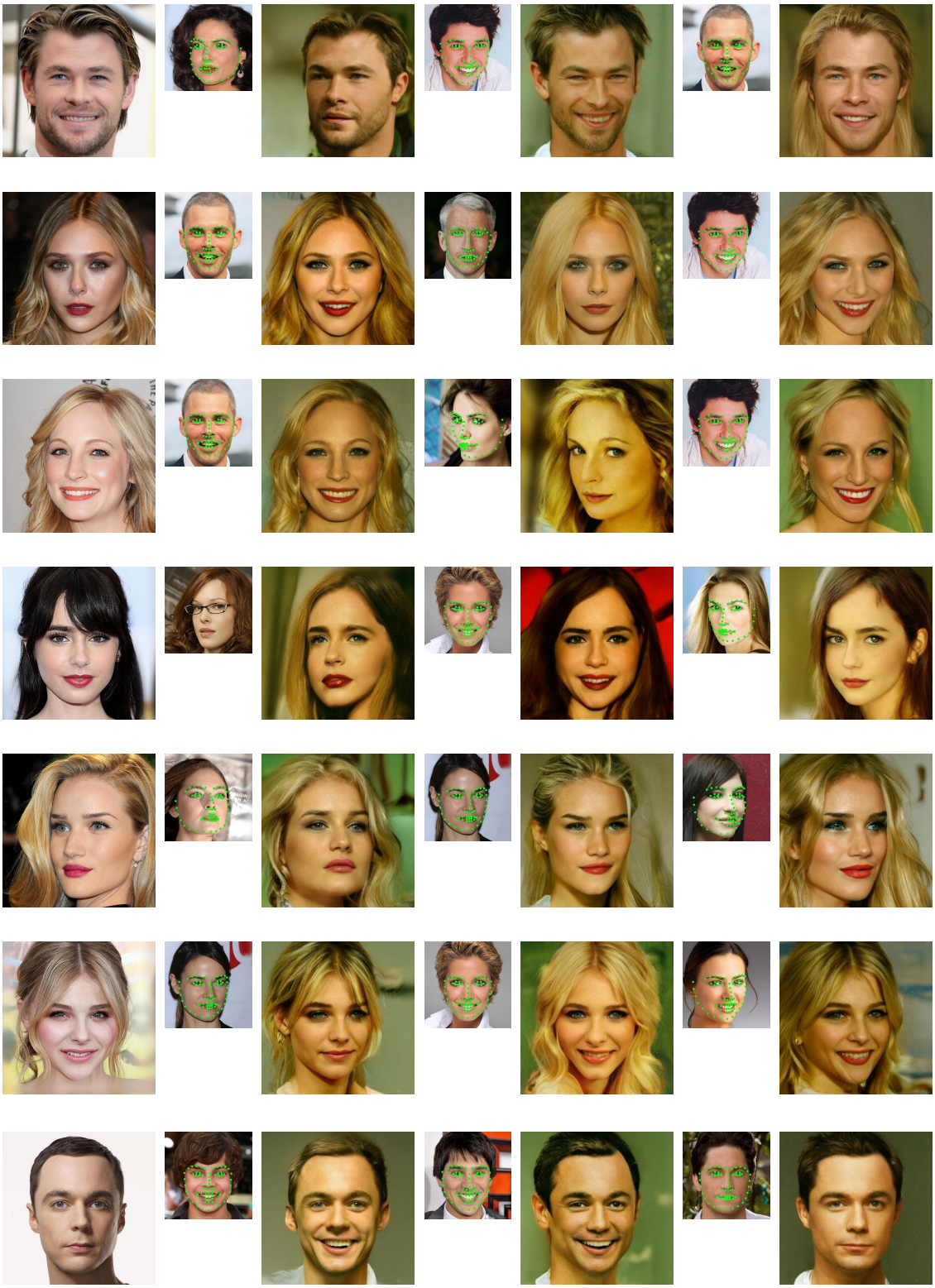


Figure 5. Generated human faces for the ID and landmark preservation task. We select three distinct landmarks to guide the generation process, incorporating variations in face orientation and facial expressions.

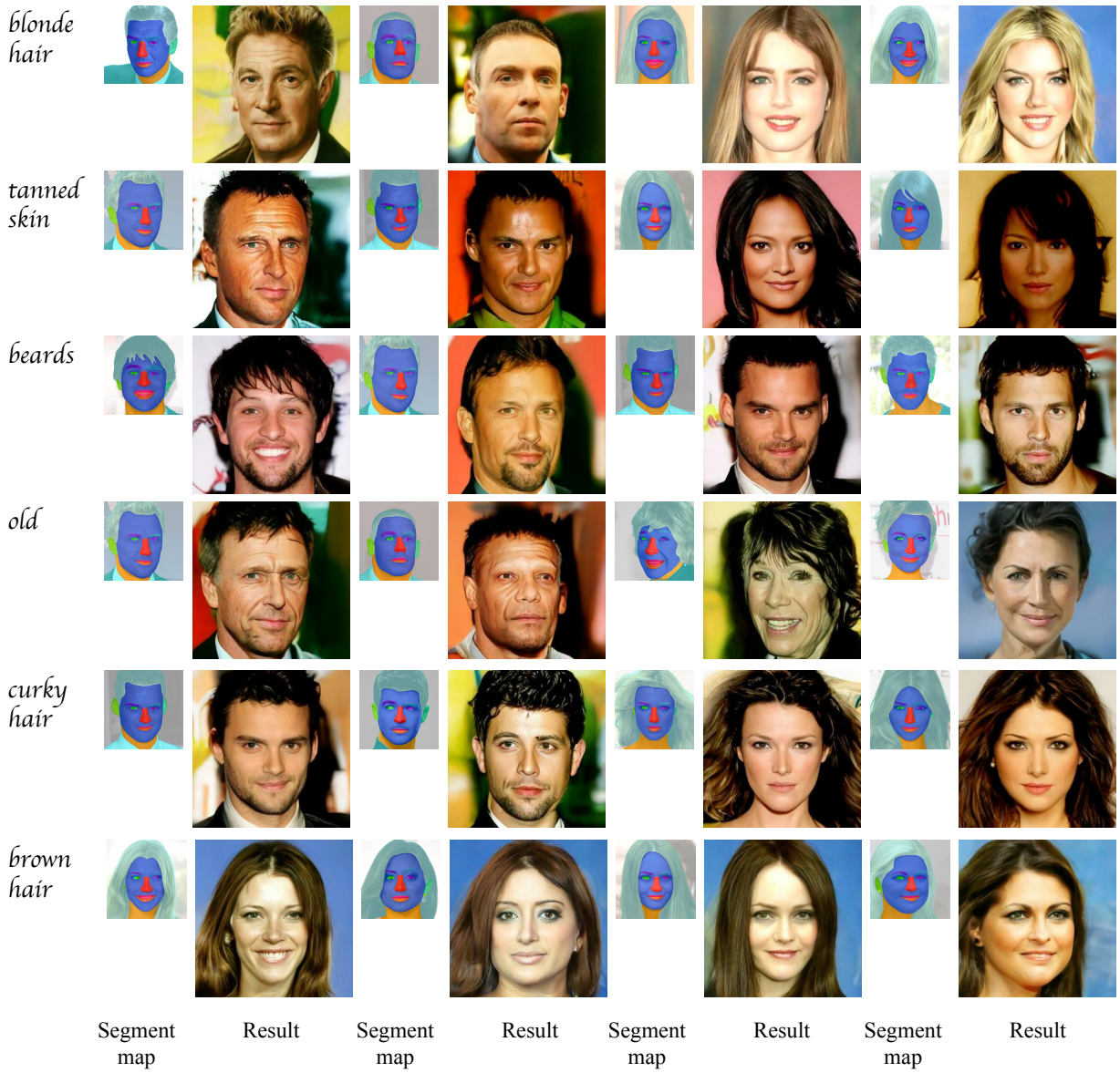


Figure 6. Generated human faces for the text and face parsing preservation task are presented. We use four distinct face segmentation maps to guide the generation process, ensuring diversity in the structural features of the output. Additionally, we incorporate variations in prompts, allowing control over one of the key attributes of the target faces.

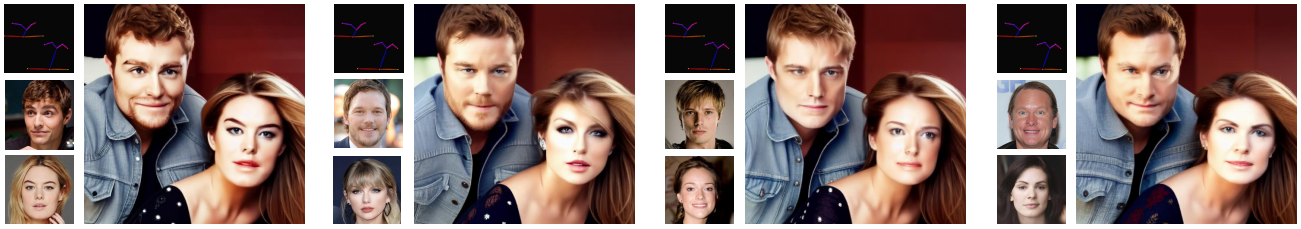


Figure 7. Generated group photos conditioned on the pose and two face identities, involving three conditions. We use different face IDs to demonstrate that our approach effectively preserves the pose prior while reconstructing the target faces.

Figure 8. Generated human faces for the ID, text, and segmentation map-guided conditional creation task. We use the same face identity combined with different prompts and face segmentation maps to demonstrate the results.

Wu, Jun-Yan Zhu, and Stefano Ermon. 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, 2022. 3

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 1

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 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, pages 10674–10685. IEEE, 2022. 3

[6] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1

[7] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 23117–23127. IEEE, 2023. 1

[8] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3813–3824. IEEE, 2023. 3