– Supplementary Material for –
Move Anything with Layered Scene Diffusion

Jiawei Ren[1,2,*] Mengmeng Xu[1] Jui-Chieh Wu[1] Ziwei Liu[2] Tao Xiang[1] Antoine Toisoul[1]

[1]Meta AI [2]S-Lab, Nanyang Technological University

{jiawei011,ziwei.liu}@ntu.edu.sg {frostxu,jerryjcw,txiang,atoisoul}@meta.com

Contents

A. Solution to Equation 10

The analytical solution to Equation 10 is:

$$f_k^{(t-1)} = \frac{\sum_n w_n \overline{\text{move}}(\alpha_k \odot \hat{v}_n^{(t-1)}, -o_{k,n})}{\sum_n w_n \overline{\text{move}}(\alpha_k, -o_{i,n})}; \quad \forall k \in \{1, \cdots, K\}, \tag{1}$$

where $n \in \{1, \cdots, N\} \cup \{a\}$, $o_{k,a}$ is the layout of the given image.

B. Discussion on Layer Masks

B.1. Elliptical blob masks

We mainly use bounding boxes for layer masks in the main paper. The layer masks can also be represented by other shapes, for example, elliptical blobs [6]. Blobs are parameterized by centroids, scales, and angles. Moreover, blobs have alpha values decaying from the centroids to soften the edges. The edge sharpness can be controlled by a parameter c: a smaller c leads to stronger edge sharpness and $c = 0$ corresponds to hard thresholding. Due to the standard Gaussian noise assumption at the initial stage of diffusion, we set $c = 0$ so that alpha values are binary. We show results of using blobs for layer masks in Figure 1.

B.2. Soft masks with modified α-blending

Soft masks can be enabled by a modified rendering equation. As discussed in the main paper, the standard Gaussian noise assumption introduced by image diffusion models requires $\sum_{k=1}^{K} \alpha_k^2 = 1$. On the other hand, the standard α-blending described in Equation 4 results in alpha values that sum to one. Therefore, the assumption can only be fulfilled when α is binary. To use soft masks, we may modify α-blending to:

$$\alpha_k = \overline{\text{move}}(m_k, o_k) \prod_{j=1}^{k-1} \sqrt{(1 - \overline{\text{move}}(m_j, o_j)^2)}, \tag{2}$$

which ensures $\sum_{k=1}^{K} \alpha_k^2 = 1$ given an all-one background. For soft masks, we use two blobs with $c = 0.05, s = 20$ and $c = 0.1, s = 10$ respectively, where s is a parameter that controls the blob size. We show results rendered by the modified α-blending in Figure 1.

*Work done during an internship at Meta AI.

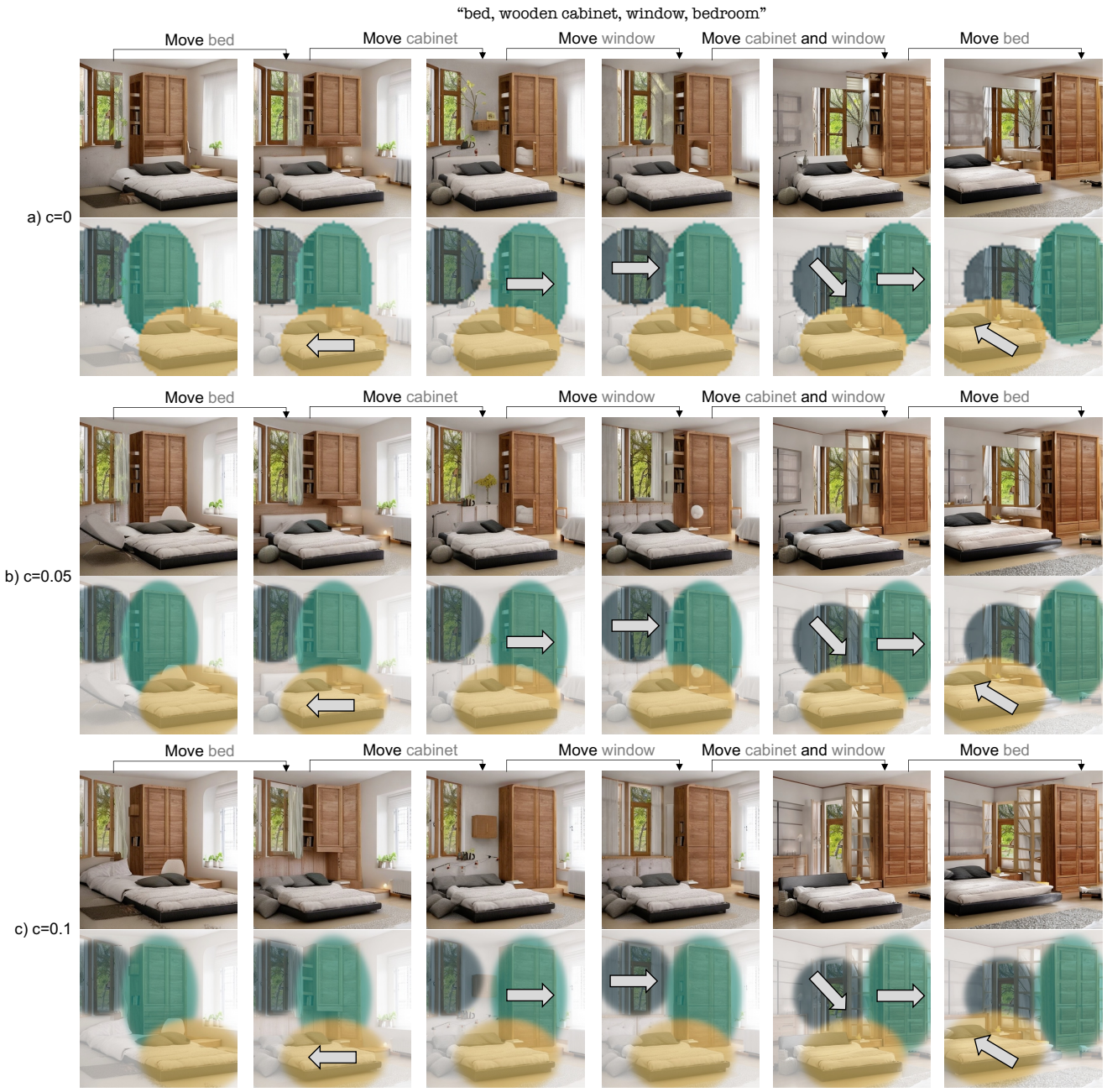

Figure 1. **Blobs as layer masks.** Layer masks can also be represented using elliptical blobs instead of bounding boxes. In addition, the updated α-blending can handle soft masks instead of binary masks.

C. Related Works

C.1. Text-to-image diffusion models

Recently, diffusion models have demonstrated unprecedented results on text-to-image generation [5, 10, 16, 19, 21], i.e., the task of generating an image from a textual description, by learning to progressively denoise an image from an input standard Gaussian noise. In the literature, T2I models vary with different design choices, including generation in pixel space [21] or latent space [19] and different denoiser architectures including U-Net [20]-based [10] or transformer [27]-based [17]. Unlike previous image editing approaches that leverage attention cues [2, 7, 8, 26] or feature correspondence [15, 23, 25], our approach is agnostic to the specific design choice of the denoiser.

C.2. Layout conditioned image diffusion

Extensive study has been made to add layout conditions to text-to-image diffusion. For training-free approaches, MultiDiffusion [1] and locally conditioned diffusion [18] predict noise using local prompts and composite them with region masking, Layout-Guidance [3] leverages the cross-attention map to provide the spatial guidance. For training-based approaches, ControlNet [29] and GLIGEN [11] finetunes the pretrained image diffusion model on paired layout-image datasets. Different from the setting in this paper, they do not focus on spatial disentanglement, thus changing layouts will also affect contents. Additionally, a line of work studies joint layout and content conditioning. Paint-by-Example [28] position reference objects to specific locations of a given image through additional model tuning, Collage Diffusion [22] harmonizes the collage of reference images using the image-to-image technique [14] improved by ControlNet [29]. Recently, a concurrent work Anydoor [4] demonstrates object moving using the paint-by-example pipeline. Our framework provides a mid-level representation and hence enables controllable scene generation, which is beyond the capability of these works.

D. Experiment Details

D.1. Dataset

Caption Generation. We use a large language model to automatically generate image captions. The prompt we used is: *Please give me 100 image captions that describe a single subject in a scene. The format is as follows: "A cat is sitting in a museum. Subject: cat. Scene: museum.". "Cat" is the subject and "museum" is the scene.* Example image captions are as follows:

1. *A bird is perched on a windowsill. Subject: bird. Scene: windowsill.*
2. *A goldfish swims in a bowl. Subject: goldfish. Scene: bowl.*
3. *A kite soars above the beach. Subject: kite. Scene: beach.*
4. *A bicycle leans against a brick wall. Subject: bicycle. Scene: brick wall.*
5. *A turtle crawls along a sandy path. Subject: turtle. Scene: sandy path.*
6. *A sunflower stands tall in a garden. Subject: sunflower. Scene: garden.*
7. *A butterfly rests on a blooming flower. Subject: butterfly. Scene: blooming flower.*
8. *A tree casts its shadow on a playground. Subject: tree. Scene: playground.*
9. *A cloud drifts over a mountain peak. Subject: cloud. Scene: mountain peak.*
10. *A snake slithers through the tall grass. Subject: snake. Scene: tall grass.*

Subject and scene descriptions are used as foreground and background local descriptions respectively. We query the language models 10 times to collect 1,000 image captions.

Image Generation. We use an open-source 512×512 text-to-image latent diffusion model to generate images from the image captions. We generate 20 images for each caption, which results in 20,000 images. Then, we use an open-vocabulary segmentation model GroundedSAM [12] to segment the foreground object. The following rule-based filters are used to remove images with no or ambiguous foreground objects:

- No bounding box detected.
- Bounding box confidence lower than 0.5.
- Bounding box area is larger than 60% of the image size.
- Segmentation mask is smaller than 5% of the image size.

5,092 images are left after filtering. Each image is associated with an image caption, local descriptions, and a segmentation mask.

D.2. Metrics

We detail evaluation metrics as follows:

- **Mask IoU.** We employ the segmentation model to predict the foreground mask on the generated images. One of the two target layouts contains the original annotated mask. We can, therefore, compute a mask IoU between the annotated mask and the shifted mask.
- **Consistency.** We compute the mask IoU between the foreground masks for the two generated images. To compensate for masks that move out of the canvas, we align the masks in two different layouts respectively and take maximum IoU.
- **Visual Consistency.** For two images generated from different layouts, we segment foreground objects out, paste them on the same location on a white canvas, and compute LPIPS to measure object-level visual consistency.
- **LPIPS.** We compute the LPIPS distance between the two generated views to examine the cross-view perceptual consistency.
- **SSIM.** We compute the SSIM similarity between the two generated views to examine the structural similarity.
- **FID.** We compute the FID between the edited images and the test dataset to evaluate the image quality.

In addition, we report KID and CLIP Score.

- **KID.** Similar to FID, we report KID as well for image quality evaluation.
- **CLIP Score.** We measure the similarity between the image embedding and the text embedding to ensure that the text alignment does not degrade after editing.

D.3. Implementation

We implement our approach on the Diffusers library using publicly available text-to-image latent diffusion models. It

employs a 64×64 latent and generates 512×512 image. For classifier-free guidance [9], we set the guidance scale to 7.5. We employ the DDIM sampler [24] and the number of sampling steps is 50. For most qualitative experiments, we set $N = 8$, $\tau = 25$, and μ_k, ν_k to 40% of the image size. For image editing experiments, we use GroundedSAM [12] to segment objects and use the segmentation masks as layer masks with manually assigned local prompts. We run all experiments on a single machine equipped with 8 32GB NVIDIA V100 GPUs. With multi-GPU parallelization, the total running time of a scene optimization and inference is less than 5 seconds.

E. Qualitative Results

E.1. More generated scenes

We show more examples of controllable scene generation in Figure 2.

E.2. Comparison of object moving

We provide a comparison with Self-Guidance [7] and a specialized inpainting model on object moving in Figure 3.

E.3. Real image editing

Our approach can edit in-the-wild images. We demonstrate multi-object moving on real images using examples provided by Epstein et al. [7] in Figure 4.

E.4. Compatibility with different denoisers

Our approach is compatible with general text-to-image diffusion models. We use a DDIM sampler and a 512×512 latent diffusion model in the main paper and show in Figure 5 that our approach also works with different samplers:

- **DPMSolver.** We set $T = 25$ and $\tau = 12$ and the inference gets even faster. We use the same random seed as the scene shown in Figure 1-Top to show the difference from DDIM-sampled results.

and different denoiser architectures:

- **An open source 1024×1024 latent diffusion model.** The model has a larger latent space and generates higher-resolution images compared to the model we used in the main paper. It also employs a different language conditioning mechanism.
- **An open source pixel diffusion model.** The model denoises on the pixel space. It has three stages, the first stage generates a 64×64 image, and the second and the third stage upsample the image to 1024×1024 resolution. Here we only show the output from the first stage.

E.5. Different random seeds

Although our approach keeps the content consistent in different views of a scene, the randomness can be introduced by changing the random noise during initialization. We show the results of three different random seeds for the object moving tasks in Figure 6.

E.6. Scenes after object replacement

A scene remains rearrangeable after object replacement. We show results of manipulating scenes with replaced objects in Figure 7.

F. Quantitative Results

F.1. Full results for controllable scene generation

We show full results for controllable scene generation with standard deviations in Table 1.

F.2. Full results for object moving comparions

We present full results for object moving comparisons with standard deviations, KID, and CLIP score in Table 2

F.3. Full results for ablation on scene generation

We show full results for N and τ ablation on controllable scene generation with standard deviations in Table 3.

F.4. Additional results for object moving ablation

We provide additional results for N and τ ablation on object moving in Table 4.

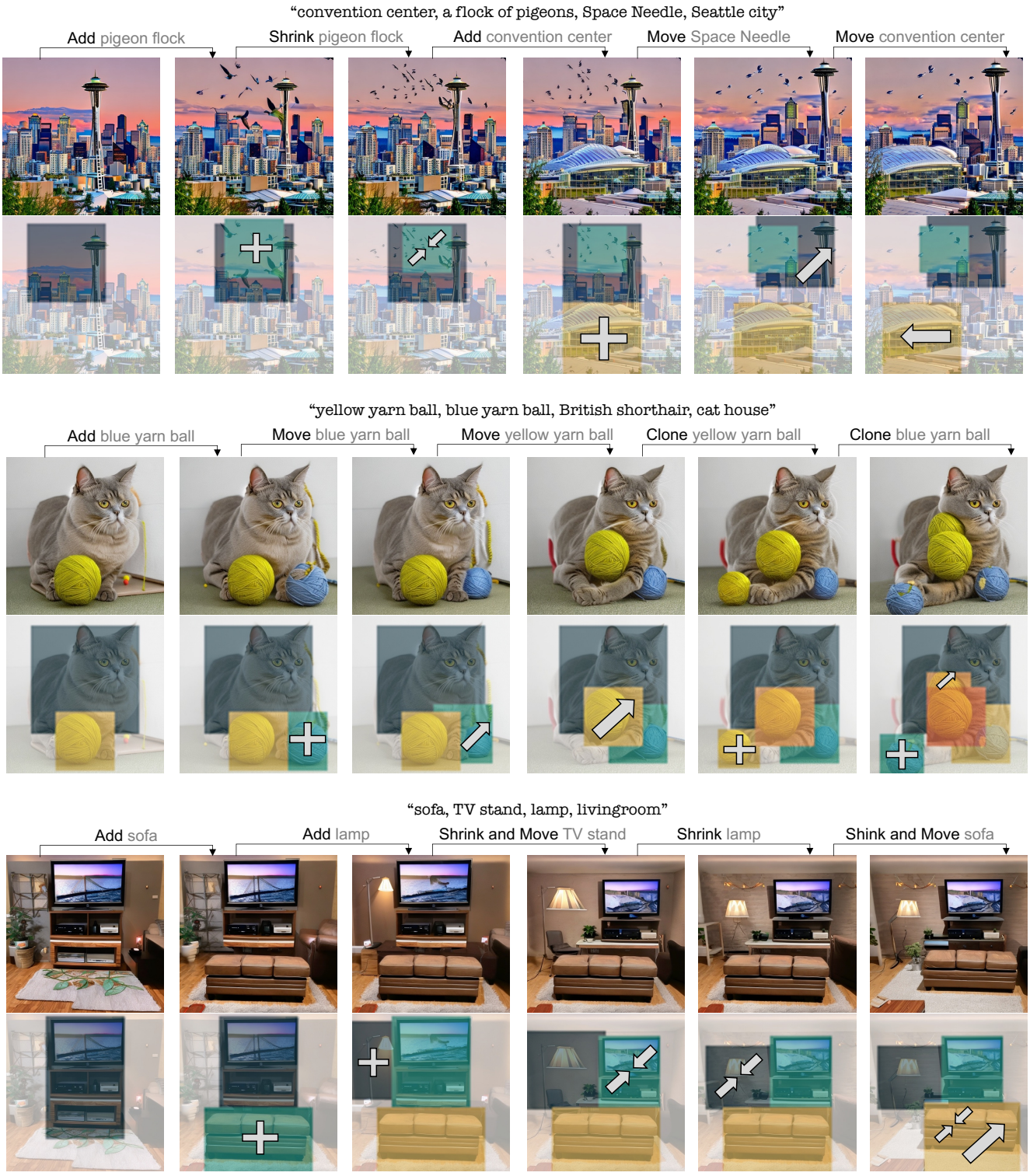

Figure 2. **More examples of generated controllable scene.** We apply sequential manipulations using the layered control.

Table 1. **Quantitative comparison for controllable scene generation.** †: without the solid color bootstrapping strategy.

Method	Mask IoU ↑	Consistency ↑	LPIPS ↓	SSIM ↑
MultiDiffusion [1]†	0.263 ± 0.004	0.257 ± 0.002	0.521 ± 0.002	0.450 ± 0.002
MultiDiffusion [1]	0.466 ± 0.001	0.436 ± 0.004	0.519 ± 0.001	0.471 ± 0.002
Ours†	0.310 ± 0.002	0.609 ± 0.003	**0.198 ± 0.001**	0.761 ± 0.001
Ours	**0.522 ± 0.001**	**0.721 ± 0.002**	0.215 ± 0.001	**0.762 ± 0.000**

Table 2. **Object moving comparison of RePaint [13], Inpainting, and our method.** †: Inpainting means a specialized inpainting model trained with masking.

Method	FID ↓	KID $_{\times 10^3}$ ↓	Mask IOU ↑	CLIP Score ↑	LPIPS ↓	SSIM ↑
RePaint	10.267 ± 0.020	1.167 ± 0.026	0.620 ± 0.001	0.321 ± 0.000	0.278 ± 0.001	0.671 ± 0.000
Inpainting†	6.383 ± 0.039	0.099 ± 0.014	0.747 ± 0.002	0.321 ± 0.000	0.264 ± 0.001	0.680 ± 0.001
Ours	**5.289 ± 0.022**	**0.059 ± 0.014**	**0.817 ± 0.003**	0.321 ± 0.000	**0.263 ± 0.001**	**0.709 ± 0.000**

Table 3. **Ablation on controllable scene generation.** We compare our method by varying the number of views N and image diffusion steps τ. †: Layout using deterministic sampling at fixed intervals.

N	τ	Mask IoU ↑	Consistency ↑	LPIPS ↓	SSIM ↑
2	25	0.477 ± 0.020	0.619 ± 0.017	0.274 ± 0.004	0.697 ± 0.004
8†	25	0.485 ± 0.006	0.638 ± 0.011	0.269 ± 0.002	0.699 ± 0.004
8	25	0.499 ± 0.005	0.657 ± 0.012	0.274 ± 0.001	0.689 ± 0.004
2	25	0.477 ± 0.020	0.619 ± 0.017	0.274 ± 0.004	0.697 ± 0.004
2	13	0.483 ± 0.024	0.661 ± 0.023	0.227 ± 0.004	0.753 ± 0.003
2	0	0.501 ± 0.015	0.699 ± 0.019	**0.208 ± 0.005**	**0.778 ± 0.004**
8	0	**0.515 ± 0.010**	**0.723 ± 0.016**	0.211 ± 0.002	0.767 ± 0.003

Table 4. **Object moving ablation.** We compare our method with inpainting-based approaches on object moving for varying number of views N and image diffusion steps τ.

N	τ	FID ↓	KID ↓	Mask IOU ↑	CLIP Score ↑	LPIPS ↓	SSIM ↑
2	25	5.918 ± 0.018	-0.020 ± 0.004	0.788 ± 0.003	0.322 ± 0.000	0.294 ± 0.001	0.672 ± 0.001
8	25	5.890 ± 0.032	-0.010 ± 0.004	0.794 ± 0.002	0.321 ± 0.000	0.289 ± 0.001	0.676 ± 0.000
2	38	7.401 ± 0.025	**-0.079 ± 0.009**	0.667 ± 0.003	0.322 ± 0.000	0.368 ± 0.001	0.598 ± 0.001
2	25	5.918 ± 0.018	-0.020 ± 0.004	0.788 ± 0.003	0.322 ± 0.000	0.294 ± 0.001	0.672 ± 0.001
2	13	**5.289 ± 0.022**	0.059 ± 0.014	0.817 ± 0.003	0.321 ± 0.000	0.263 ± 0.001	0.709 ± 0.000
2	0	5.320 ± 0.029	0.182 ± 0.020	**0.836 ± 0.003**	0.322 ± 0.000	**0.255± 0.001**	**0.722 ± 0.001**

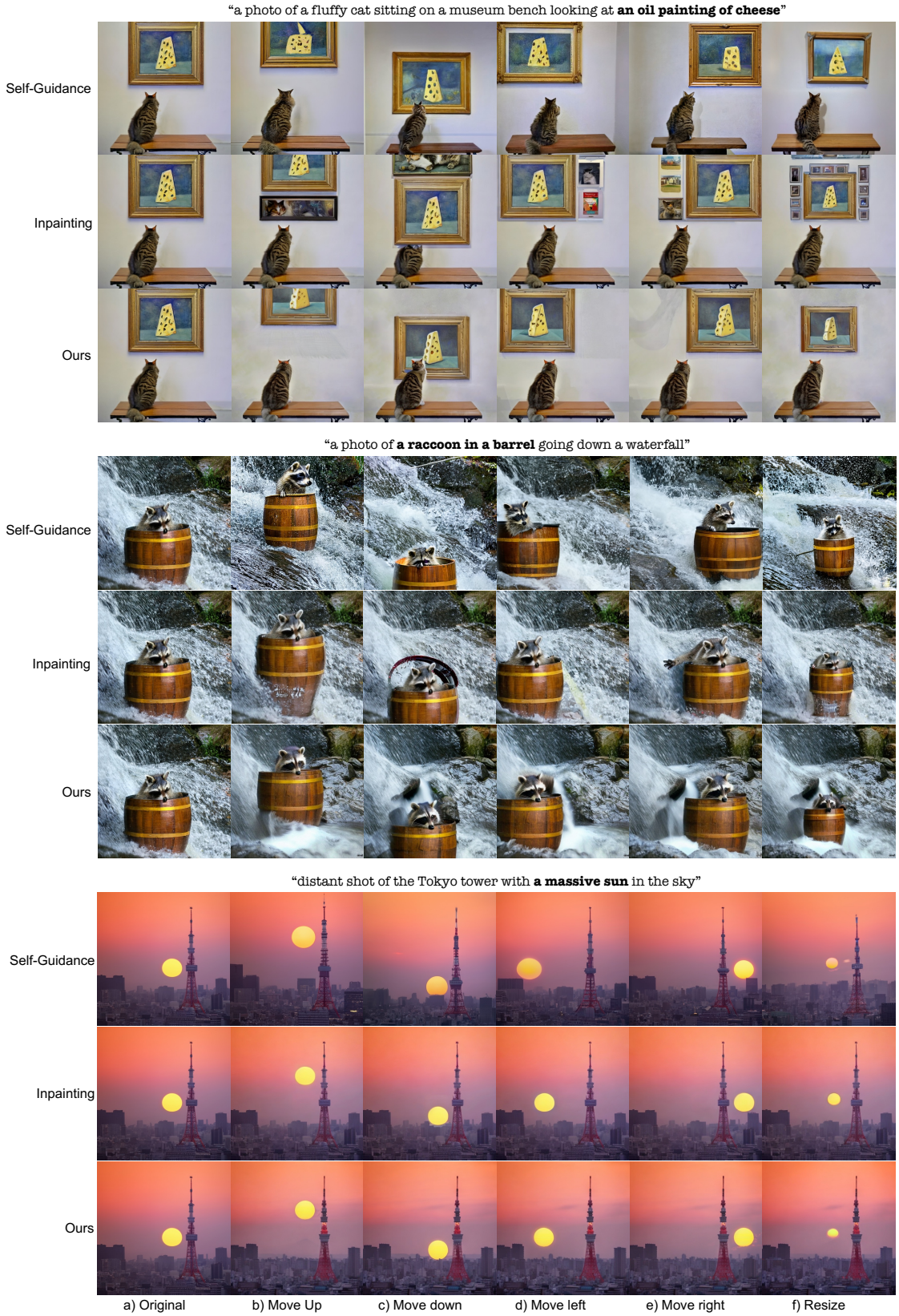

Figure 3. **Qualitative comparison on object moving.** Self-Guidance [7] and inpainting generates varing content across editings.

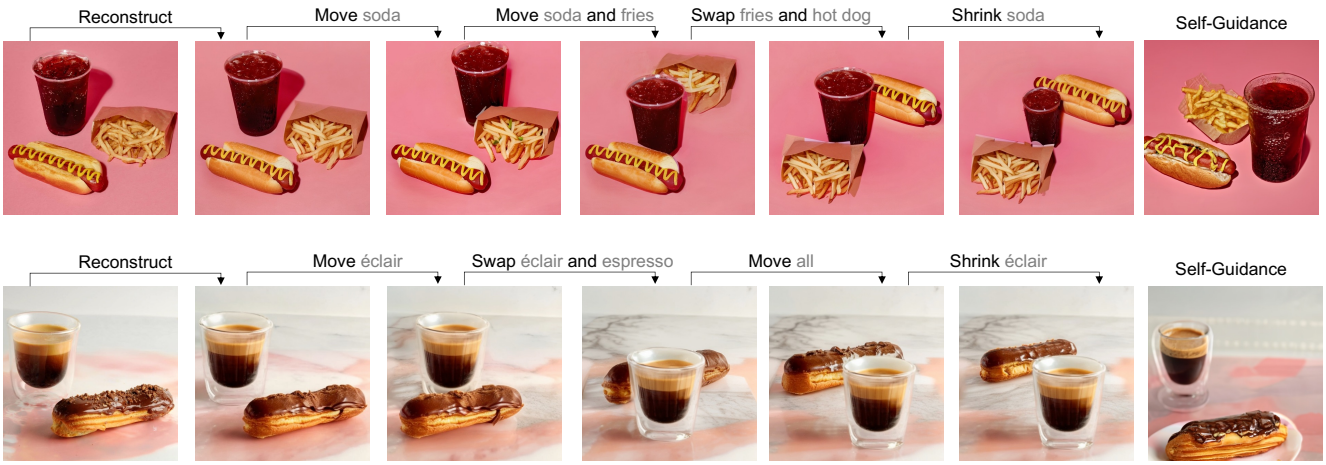

Figure 4. **Multi-object moving on real images.** Examples are borrowed from Epstein et al. [7].

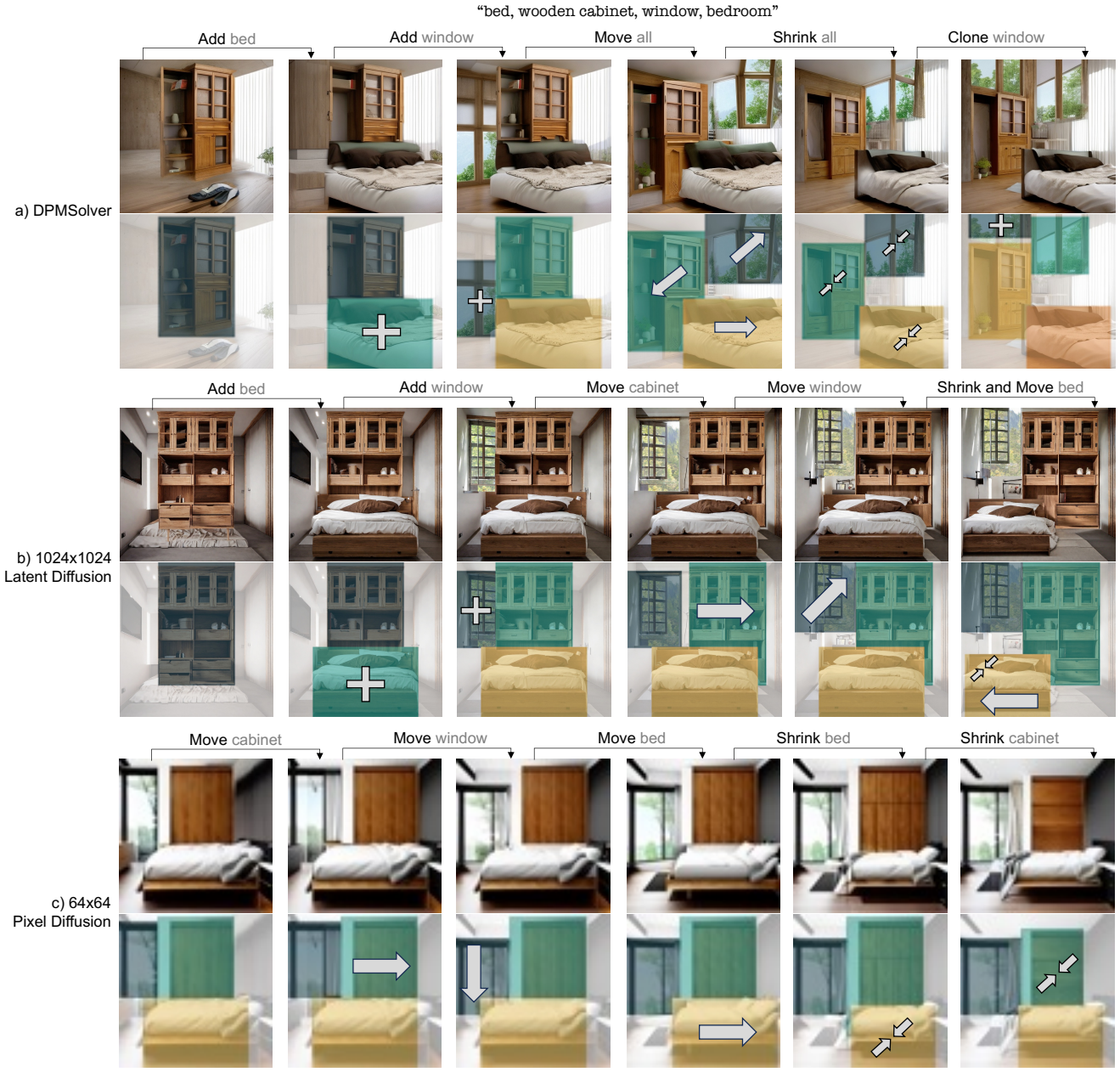

Figure 5. **Diffusion sampler and architecture.** We present editing results with different diffusion samplers and denoiser architectures to show our method is applicable in various configurations.

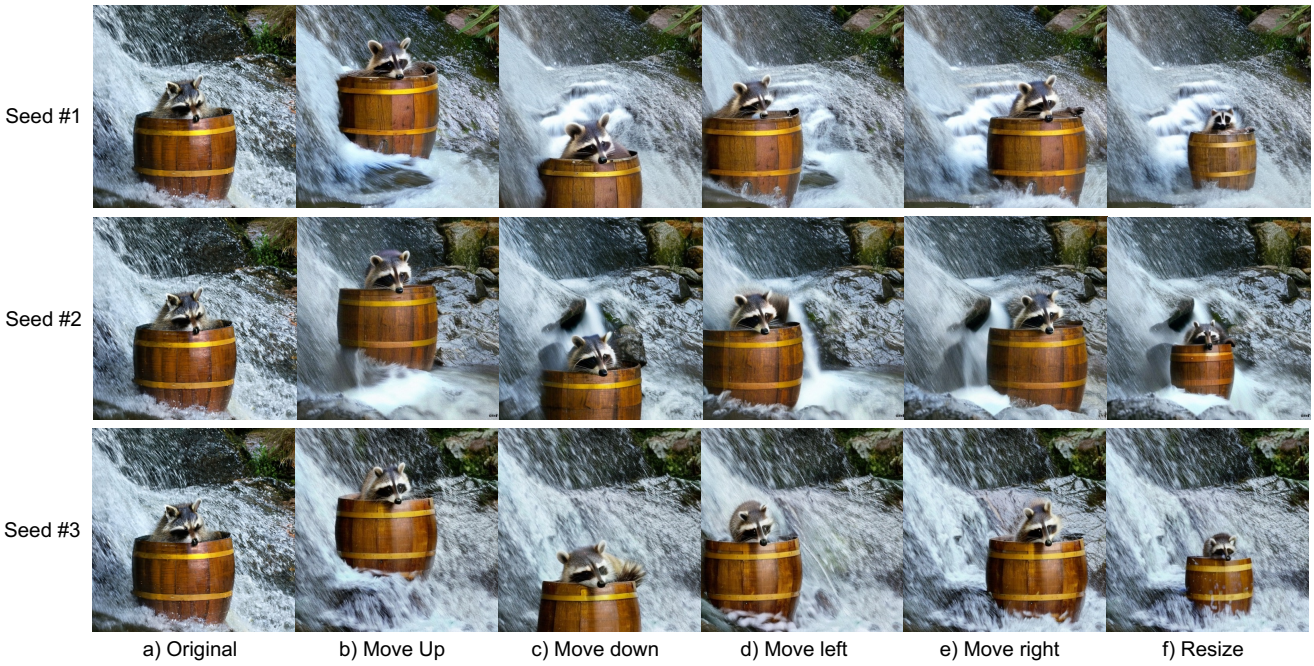

Figure 6. Results with different random seeds in the object moving task.

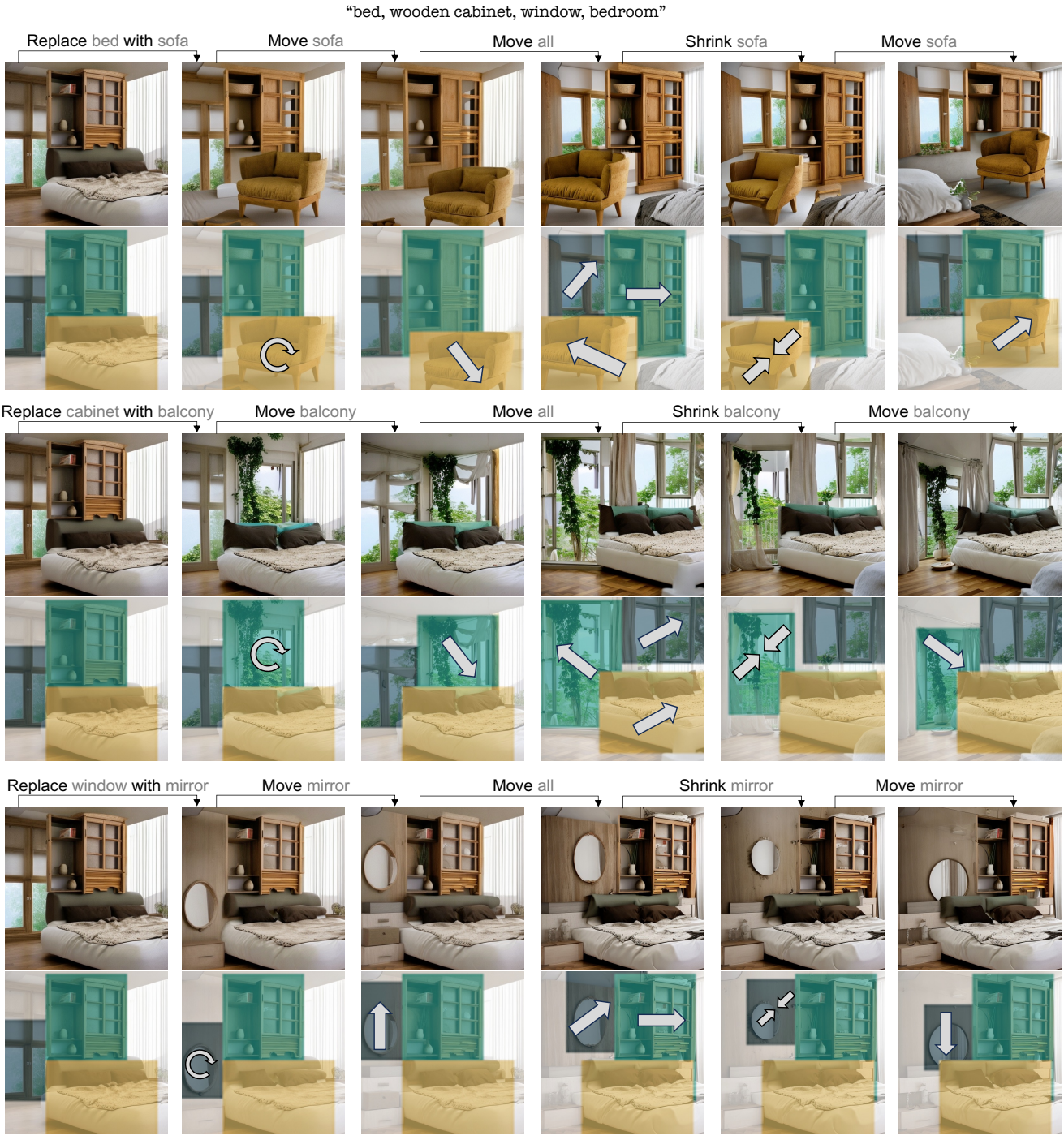

Figure 7. **Manipulating scenes with replaced objects.** We first replace an object in the scene before manipulating the scene layout show the corresponding editing results.

References

[1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *Proceedings of the 23rd International Conference on Machine Learning*, 2023. 3, 6

[2] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10, 2023. 2

[3] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. *arXiv preprint arXiv:2304.03373*, 2023. 3

[4] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*, 2023. 3

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[6] Dave Epstein, Taesung Park, Richard Zhang, Eli Shechtman, and Alexei A Efros. Blobgan: Spatially disentangled scene representations. In *European Conference on Computer Vision*, pages 616–635. Springer, 2022. 1

[7] Dave Epstein, Allan Jabri, Ben Poole, Alexei A Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *arXiv preprint arXiv:2306.00986*, 2023. 2, 4, 7, 8

[8] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

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

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[11] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 3

[12] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3, 4

[13] 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*, pages 11461–11471, 2022. 6

[14] 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

[15] Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. *arXiv preprint arXiv:2307.02421*, 2023. 2

[16] Alex 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. *arXiv preprint arXiv:2112.10741*, 2021. 2

[17] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 2

[18] Ryan Po and Gordon Wetzstein. Compositional 3d scene generation using locally conditioned diffusion. *arXiv preprint arXiv:2303.12218*, 2023. 3

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. arxiv. *arXiv preprint arXiv:2112.10752*, 2021. 2

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[21] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 2

[22] Vishnu Sarukkai, Linden Li, Arden Ma, Christopher Ré, and Kayvon Fatahalian. Collage diffusion. *arXiv preprint arXiv:2303.00262*, 2023. 3

[23] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023. 2

[24] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4

[25] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 2

[26] 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

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[28] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 3

[29] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 3