Multi-view Image Diffusion via Coordinate Noise and Fourier Attention (Supplementary Material)

Justin Theiss Norman Müller Daeil Kim Aayush Prakash
Meta Reality Labs

A. Noise Initialization

A.1. Implementation Details

In this section we provide further implementation details of our coordinate-based noise initialization. For each set of multi-view images, we first sample a "shared noise" that is used across all views (*i.e.* ϵ_{shared} in Eqn. 7). To provide the model with low spatial frequency information related to the change in camera pose across views, we transform normalized pixel coordinates from each view into the space of the center view. We then take the cosine of these values to remap pixel coordinates into the range $[-1, 1]$. These transformed pixel coordinates are then combined with the shared noise according to Eqn. 8. The coordinate noise for each view $\hat{\epsilon}^i$ is then combined with per-view independent noise ϵ^i as shown in Eqn. 9.

A.2. Quantitative Comparisons of Noise Initialization Methods

In order to further evaluate the choice of coordinate noise, we compare against other relevant methods for incorporating shared noise or low-frequency information (Table S1). The first comparison of interest is "mixed noise" [6], which uses a combination of shared noise across views and independent noise per view. This is similar to our "shared noise" condition in our ablation study in the main paper (Table 4) but uses a different weighting scheme (Eqn. S1 with $\alpha = 1$). As shown in Table S1, our shared noise implementation provides better performance across all metrics except Intra-LPIPS (compare first two rows).

$$\epsilon^i_{\text{mixed}} = \epsilon_{\text{shared}} \frac{\alpha^2}{1+\alpha^2} + \epsilon^i \frac{1}{1+\alpha^2} \tag{S1}$$

Next, we compare the effect of using our "coordinate noise" implementation *vs*. combining low-frequency coordinate noise and high-frequency independent noise, which has been suggested in recent work conditioning on images [20, 27]. Although we do not condition directly on image frames, it's clear that the combination of low-frequency coordinate noise and high-frequency independent noise is not as effective as our implementation using Eqn. 9 (compare last two rows of Table S1).

Overall, it is interesting to note that although our coordinate noise method provides substantial improvements in FID and overlapping PSNR, mixed noise obtains better performance when measuring Intra-LPIPS.

Table S1. Comparison of noise initialization methods in the panoramic experiment.

Method	FID ↓	CLIP Score ↑	PSNR ↑	Ratio ↑	Intra-LPIPS ↓
Mixed Noise [6]	23.25	24.69	23.25	0.624	**0.719**
Shared Noise (Eqn. 7)	22.06	24.71	23.63	0.635	0.794
Low Freq. Coord. Noise	36.99	23.14	21.63	0.582	0.777
Coord. Noise (Eqn. 8)	**19.55**	**24.95**	**24.25**	**0.651**	0.776

B. FID/CLIP Score Differences Between Experiments

As noted in the main paper, we observed improved performance as measured by FID and CLIP Score compared to MVDiffusion in the panoramic but not the depth-to-image experiment (*cf*. Tables 1 & 3). One explanation for this performance difference is that ScanNet text prompts provided by [24] using blip2 were often imprecise or inconsistent across views. Since MVDiffusion's method does not account for non-overlapping regions, their method is susceptible to issues like that shown in Figure S1 for imprecise prompts (here, the prompt "a pair of shoes sitting on the floor next to a bed" leads to hallucinations of a second bed). These errors can lead to better CLIP Score performance at the expense of multi-view consistency. Furthermore, inconsistent prompts across a scene could negatively impact FID for our method compared with MVDiffusion, which may exhibit errors only in single views without reconciling across a scene.

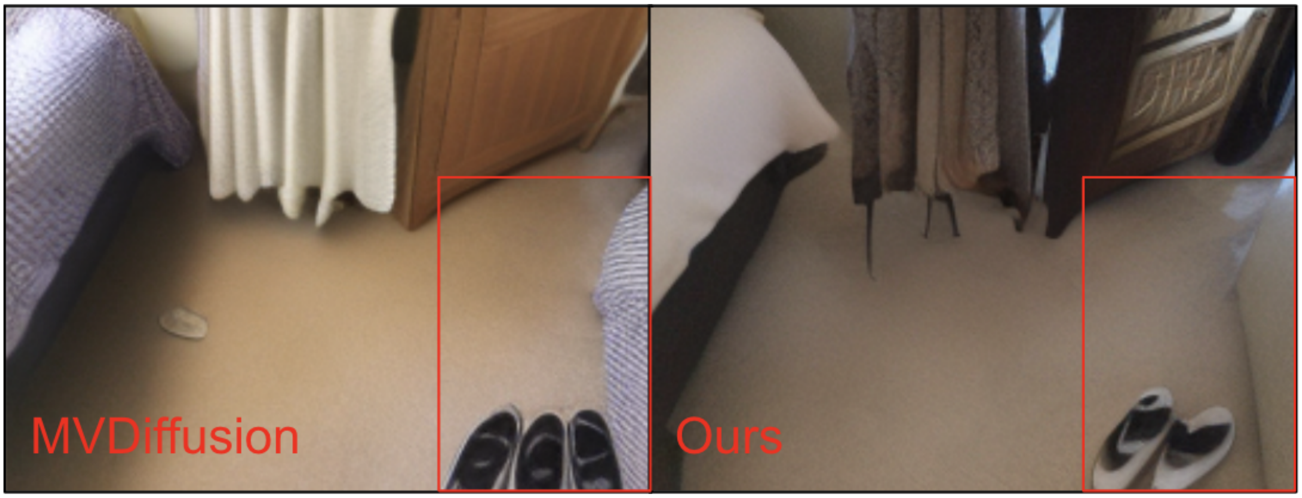

Figure S1. MVDiffusion *vs*. our method with an imprecise prompt "*a pair of shoes sitting on the floor next to a bed*."

C. Additional Ablation Studies

In order to further evaluate our design choices for noise initialization, we compare results from experiments varying the weight parameter w from Eqn. 8. The results shown in Table S2 indicate that setting the weight $w = 0.5$ indeed provides the optimal result. However, it is interesting to note that this paramter appears to primarily affect FID and overlapping PSNR metrics. For these metrics, performance is noticeably – albeit not substantially – worse in either direction away from 0.5.

Table S2. Ablation of weight parameter w in Eqn. 8 in the panoramic experiment.

Method	FID ↓	CLIP Score ↑	PSNR ↑	Ratio ↑	Intra-LPIPS ↓
Shared Noise ($w = 0.0$)	22.06	24.71	23.63	0.635	0.794
Coord. Noise ($w = 0.25$)	19.71	24.90	23.92	0.643	0.779
Coord. Noise ($w = 0.5$)	**19.55**	**24.95**	**24.25**	**0.651**	**0.776**
Coord. Noise ($w = 0.75$)	21.02	24.90	23.59	0.634	0.787
Coord. Noise ($w = 1.0$)	21.70	24.90	23.91	0.643	0.781

We additionally compare performance when using a binary high pass filter (HPF) mask (Eqn. 13) *vs*. a Gaussian HPF approach as well as when using a time-dependent (HPF-r_t) *vs*. constant low pass stop frequency (LPF-0.25, using stop frequency from [20, 27]). The results shown in Table S3 demonstrate that there is minimal difference between the binary or Gaussian HPF mask. However, we observe that using a time-dependent HPF mask provides substantially better performance.

Table S3. Comparison of binary and Gaussian high (HPF) or low (LPF) pass filters (Eqn. 13) in the panoramic experiment.

Method	FID ↓	CLIP Score ↑	PSNR ↑	Ratio ↑	Intra-LPIPS ↓
Gaussian LPF-0.25 mask	23.99	24.71	23.13	0.621	0.771
Gaussian HPF-r_t mask	22.59	**24.84**	24.47	0.657	0.762
Binary HPF-r_t mask (Eqn. 13)	**22.36**	24.68	**24.67**	**0.662**	**0.755**

Note: Filters are either time-dependent (*i.e*. "HPF-r_t" where r_t is the radius defined in Eqn. 13) or use a normalized stop frequency of 0.25 (*i.e*. "LPF-0.25").

Finally, we further validate the design choice of our time-dependent Fourier-based attention module. Specifically, we consider the following conditions: no spatial frequency filtering ("No filter"), time-dependent low pass filtering ("LPF-r_t"), as well as low and high pass filtering using the inverse relationship with denoising time steps ("LPF-$(1 - r_t)$" and "HPF-$(1 - r_t)$", respectively). In the latter two conditions, the radius r_t of the spatial frequency mask in Eqn. 13 decreases from 1 to 0 across denoising time steps. For low pass filtering (*i.e.* "LPF-$(1 - r_t)$"), this means that all frequencies are included in $\bar{\mathbf{G}}_t^j$ (Eqn. 14) at the noisiest time steps and only the lowest frequencies are included at the least noisy time steps.

As shown in Table S4, our method of selecting the full spectrum of spatial frequencies for attention at noisier time steps and high spatial frequencies at less noisy time steps (*i.e.* "HPF-r_t") provides the best overall performance, particularly for FID and overlapping PSNR. Similar to our ablation of the weight parameter in Eqn. 8 (Table S2), we observe relatively less variation across conditions for the CLIP Score and Intra-LPIPS metrics.

Table S4. Comparison of time-dependent low or high pass filters in the panoramic experiment.

Method	FID ↓	CLIP Score ↑	PSNR ↑	Ratio ↑	Intra-LPIPS ↓
No filter	25.89	**24.85**	23.31	0.626	0.747
LPF-$(1 - r_t)$	29.71	24.78	22.66	0.609	0.768
LPF-r_t	23.81	24.75	24.12	0.648	**0.740**
HPF-$(1 - r_t)$	23.57	24.57	24.00	0.645	0.772
HPF-r_t (Eqn. 13)	**22.36**	24.68	**24.67**	**0.662**	0.755

Note: The low pass filter (LPF) is defined as $1 - \mathbf{M}_{\mathcal{F}}^{r_t}$ and, *e.g.*, "HPF-$(1 - r_t)$" implies $\mathbf{M}_{\mathcal{F}}^{(1-r_t)}$.

D. Additional Qualitative Examples

In this section, we provide further qualitative examples of our method in comparison to baselines in the depth-to-image and panoramic image generation experiments.

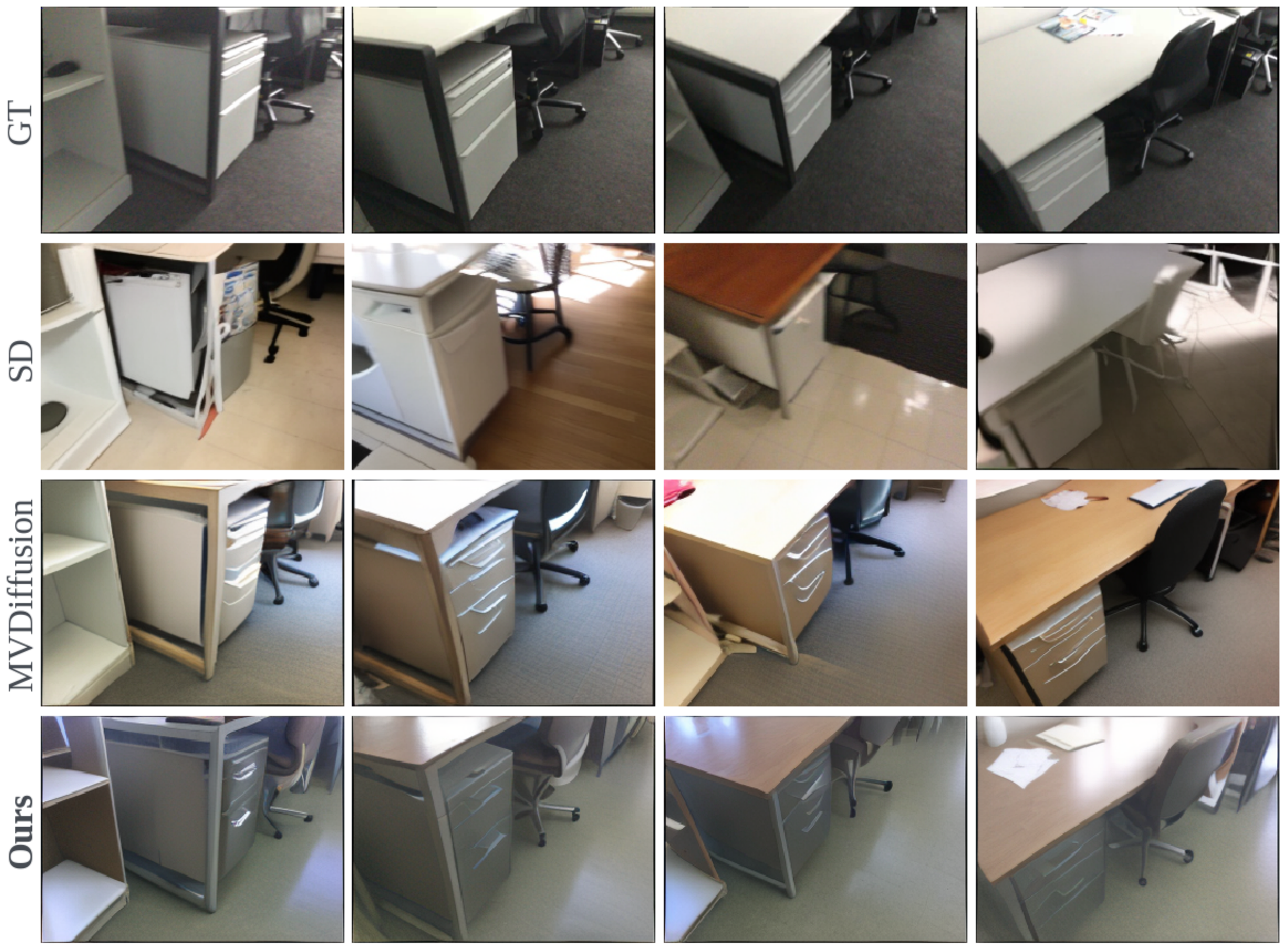

Figure S2. Depth-to-image generation using the prompt “*a desk with a chair and a filing cabinet*.”

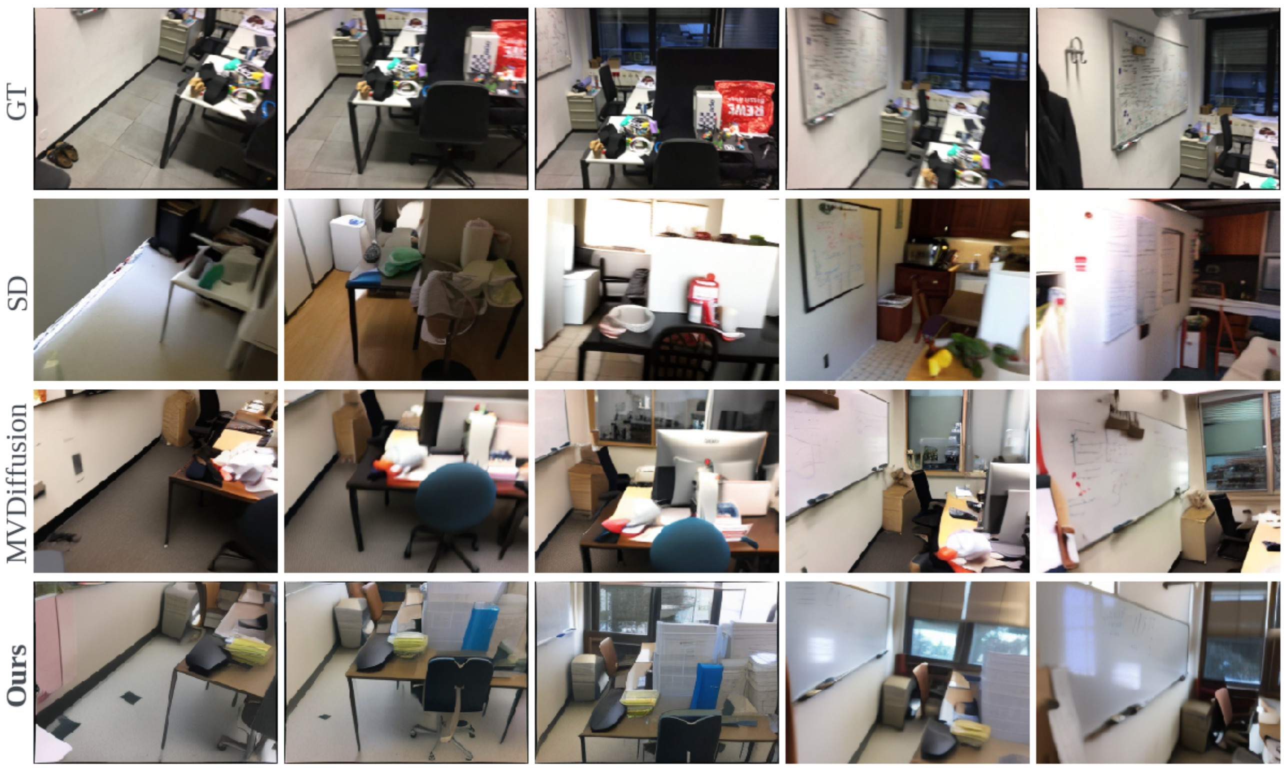

Figure S3. Depth-to-image generation using the prompt "*a whiteboard on a wall in an office*."

Figure S4. Panoramic image generation using the prompt "*a kitchen with a large black vase on the counter and a marble counter top next to a sink*."

Figure S5. Panoramic image generation using the prompt "*a living room filled with furniture and a piano*."

Figure S6. Panoramic image generation using the prompt "*a white building with a door and some plants in front of a white house with a large glass door*."

References

[1] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023. 1, 2

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 1

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 6

[4] 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), 2023. 2

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6

[6] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023. 1, 3, 4, 9

[7] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 1, 3

[8] Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549*, 2023. 1

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[10] 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. *Advances in neural information processing systems*, 30, 2017. 6

[11] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. *arXiv preprint arXiv:2303.11989*, 2023. 1

[12] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. *arXiv preprint arXiv:2306.05178*, 2023. 1, 2, 6, 7

[13] Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5404–5411, 2024. 1, 3

[14] 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. 1

[15] 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 *International Conference on Machine Learning*. PMLR, 2022. 2

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

[17] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 1, 3

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[19] 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. 1, 2

[20] Weiming Ren, Harry Yang, Ge Zhang, Cong Wei, Xinrun Du, Stephen Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024. 1, 3, 4, 9, 10

[21] 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. 1, 2, 6, 7

[22] 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. 3

[23] 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, 2022. 1, 2

[24] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv preprint arXiv:2307.01097*, 2023. 1, 3, 4, 5, 6, 7, 9

[25] 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. 5

[26] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022. 6

[27] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. *arXiv preprint arXiv:2312.07537*, 2023. 1, 3, 4, 5, 9, 10

[28] Han Xue, Zhiwu Huang, Qianru Sun, Li Song, and Wenjun Zhang. Freestyle layout-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[29] Jeffrey Zhang, Shao-Yu Chang, Kedan Li, and David Forsyth. Preserving image properties through initializations in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5242–5250, 2024. 3

[30] 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. 1, 7

[31] Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel generation of large content with diffusion models. *arXiv preprint arXiv:2303.17076*, 2023. 2

[32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6