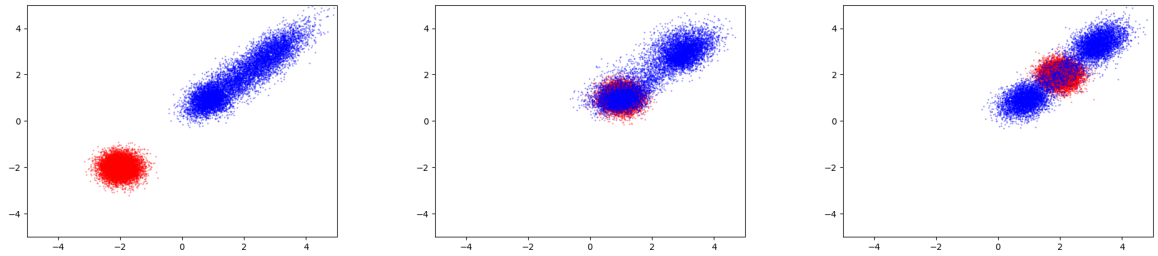
Deterministic Diffusion for Sequential Tasks: Rebuttal Additions

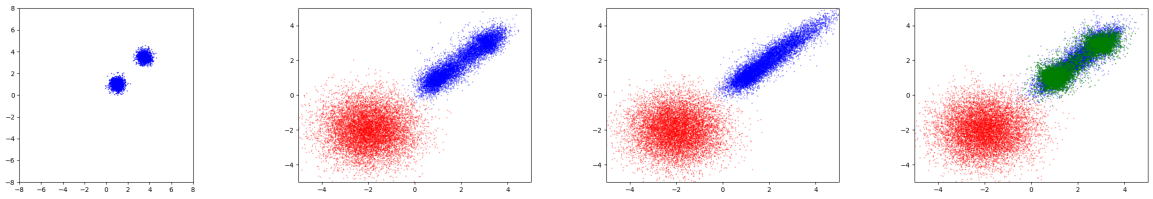
A Rectified Flow

We ran the three instructive examples described in the paper in section 4.2 based on the Rectified Flow algorithm (Liu et al., 2022) instead of Iterative α-(de)blending (Heitz et al., 2023): Experiment 1 (Fig.1), Experiment 2 (Fig.2) and Experiment 3 (Fig.3). Additional videos of the results are available on the anonymous project website. The results show that correct selection of the initial distribution is advantageous for improved performance, regardless of whether the underlying algorithm is Rectified Flow or Iterative α-(de)blending. In the videos, it can be seen that the results using Rectified Flow move in straighter lines compared to Iterative α-(de)blending results, though the resulting distributions seem similar. These results are based on three reflow operations ($N = 3$) (see Liu et al. (2022) for details).



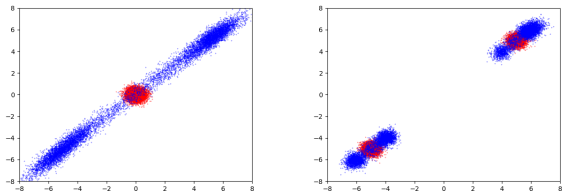
(a) $\mathcal{N}\left((-2,-2)^T, \boldsymbol{\sigma}\right)$ (b) $\mathcal{N}\left((1,1)^T, \boldsymbol{\sigma}\right)$ (c) $\mathcal{N}\left((2,2)^T, \boldsymbol{\sigma}\right)$

Figure 1: **Effects of Rectified Flow source distribution.** Source distribution for each example is described in its caption. The target distribution in all three examples is a bi-modal Gaussian, centered at $(1,1)$ and $(3,3)$. Source distribution variance is $\boldsymbol{\sigma} = 0.1\boldsymbol{I}$. Red dots are samples from the source, blue dots are samples from the model after 10 steps of the Rectified Flow process.



(a) Target (b) 10 steps (c) 3 steps (d) 3 + 7 steps

Figure 2: **Rectified Flow multi-stage blending.** Beginning with a source distribution of $\mathcal{N}\left((-2,-2)^T, \boldsymbol{I}\right)$ (depicted by red dots), Fig. 2b and Fig. 2c show results of the Rectified Flow algorithm with 10 and 3 steps respectively of sampling towards the target bi-modal Gaussian at $(1,1)$ and $(3.5,3.5)$ (shown in blue in Fig. 2a). Samples from this first model are shown in blue. Fig. 2d shows samples from the first model after 3 steps of Rectified Flow (blue), and additionally shows (in green) samples from a second model, trained to start the Rectified Flow process from the output of the first model towards the target.



(a) Unconditional source distribution (b) Conditional source distribution

Figure 3: **Effects of conditioning with Rectified Flow.** Samples from the source (red) and targeet (blue) distributions.

B Video prediction

Below are additional ablations on the PHYRE dataset as suggested by the reviewers. Note that we are running experiments of the same ablation on the BAIR dataset as well; however, due to the time constraints of the rebuttal period, results will only be available for the final version of the paper.

B.1 Ablation on PHYRE

Fig. 4 is similar to the FVD plot for the PHYRE dataset in the paper (Sec. 5, Fig. 4(b)) with the addition of experiments where the source distribution is initialized from DDIM instead of the previous frame. This initialization is similar to the $DDIM_n$ initialization used for the Diffusion Policy experiments. We experimented with $DDIM_2$, $DDIM_3$ and $DDIM_4$ initializations, in each case following up with deblending steps to complete a total of 5 or 10 steps.

The experiments conducted are as follows:

- 2 DDIM steps (Blue) followed by 8 or 3 deblending steps, bringing the total to 10 or 5 steps respectively.
- 3 DDIM steps (Orange) followed by 7 or 2 deblending steps, bringing the total to 10 or 5 steps respectively.
- 4 DDIM steps (Pink) followed by 6, bringing the total to 10 steps

On the plot in Fig. 4, each result is represented by its combined total number of steps (5 or 10).

$DDIM_3$ produced better results than $DDIM_2$ for both 10 and 5 total steps. We believe this is due to the better initial prediction; although $DDIM_4$ had the worst results, we believe this is due to not enough deblending steps. However, the $DDIM_n$ initialization did not improve results on the PHYRE dataset compared to the previously presented history-based initialization. As we hypothesize in the paper, this occurs due to the nature of the task: in video prediction frames are predicted one frame at a time, and the previous frame is a good approximation of the next frame.

C Robotic Control

In the following subsections, we present additional baselines and ablations for the robotic control domains presented in the paper. We will replace the original figures in the paper with the new ones presented here for the updated version of the paper.

C.1 Ablation on Push-T

We performed an ablation study on Push-T similar to the one done on the Tool-Hang task, presented in Fig. 10 of the paper. Results of this new study are in Fig. 5.

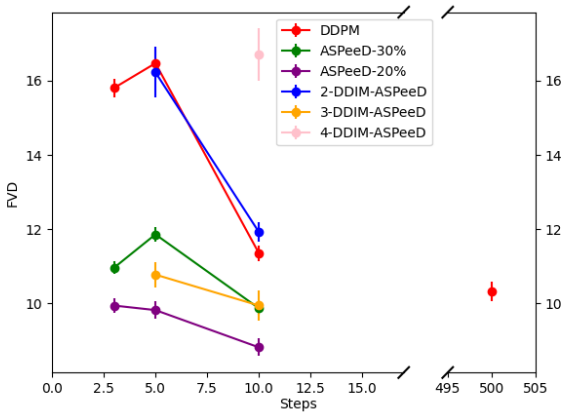


Figure 4: PHYRE FVD (lower is better) with additional results of $DDIM_n$ initialization. In blue the initialization is $DDIM_2$ followed by 3 or 8 deblending steps, in orange the initialization is $DDIM_3$ followed by 2 or 7 deblending steps, in pink the initialization is $DDIM_4$ followed by 6 deblending steps. History-based initialization outperforms $DDIM_n$ initialization, with the correct amount of perturbation noise applied.

As in the Tool-Hang task, we find it beneficial not to add any perturbation to the $DDIM_n$ source distribution, both in terms of reward and MSE. The reward is not sensitive to changing the ratio of DDIM to deblending steps, unlike the MSE.

C.2 Additional Baselines

In Fig.6 we compare our algorithm with DPM-Solver++ (Lu et al., 2022a) of order 2 and 3 with the multistep solver and in Fig.7 we add a comparison with a Consistency Model (Song et al., 2023) baseline with 10 sampling steps (we note that 10 sampling steps of the Consistency Model baseline performs slightly better than 1 sampling step). In all runs we used the DPM-Solver++ parameters suggested in the official Github repository. We implemented Consistency Models on top of the Diffusion Policy code according to Algorithm 3 (CT) described in their paper and Github repository, using L2 loss which we used for all other runs as well.

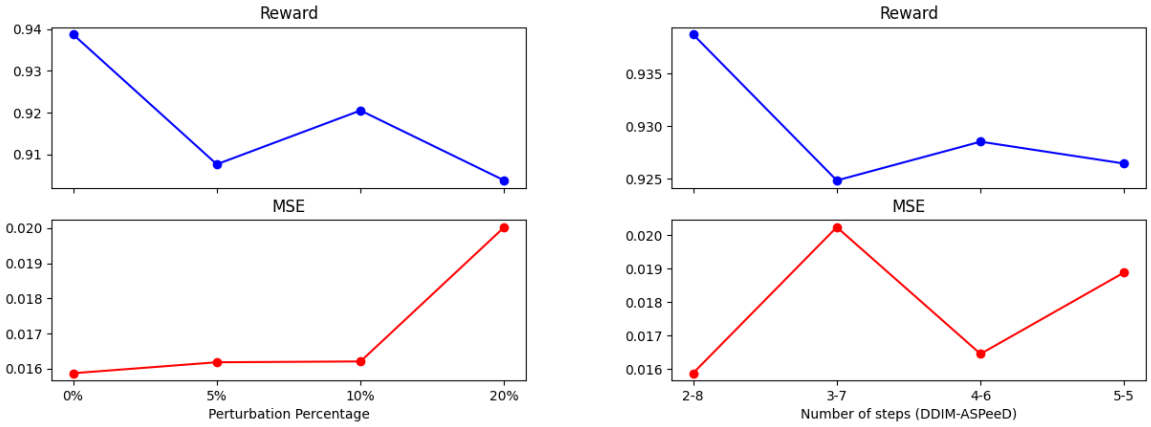
In the Tool-Hang task ASPeeD was run with 3 DDIM steps for initialization and 7 deblending steps, while in the Push-T task ASPeeD was run with 2 DDIM steps for initialization and 8 deblending steps. In total, all algorithms use 10 steps, except 100 DDPM which uses 100. Our approach is superior to all other baselines both in reward and MSE.

C.3 Push-T History-Based Initialization Ablation

We added an ablation study using the history-based initialization technique in the Push-T task presented in Fig. 8, where we duplicated the current observation and added Gaussian perturbation. We compared history-based initialization with 10 deblending steps and different levels of perturbation: $20\%, 30\%$ and 40% with the $DDIM_2$ based initialization, where 2 steps of DDIM are taken followed by 8 deblending steps. The results show that in this domain $DDIM_n$ initialization is more beneficial than history based initialization as the rewards are higher and the MSE is lower. We suspect that this is due to the nature of the prediction task: the output of the network is 16 future states, so initializing from the current state is not a good approximation of the required sample.

D Extended Related Work

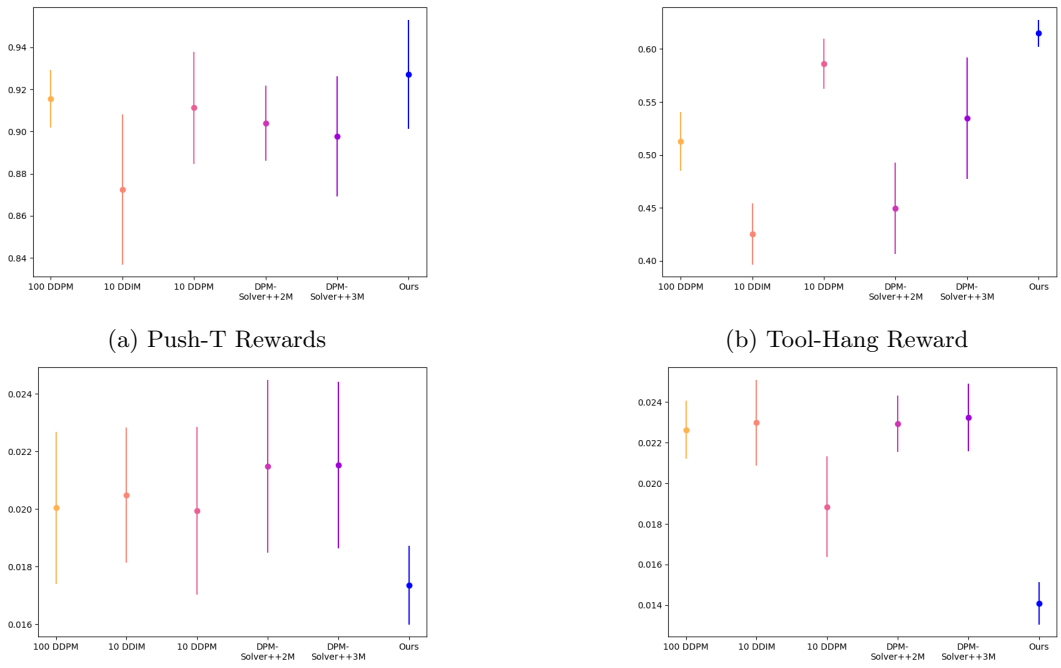
Diffusion models have exploded in popularity in recent years, and have been used in a wide variety of domains as powerful generative models, most prominently in image generation (Ho et al., 2020; Dhariwal & Nichol,



(a) Percentage of noise perturbation

(b) Number of DDIM / deterministic steps

Figure 5: **Robot control ablations.** Rewards (top) and MSE (bottom) for both our ablation experiments using a $DDIM_n$ source distribution on the Push-T domain. Fig. 5a: different variance of added Gaussian noise. Fig. 5b: Different numbers of DDIM steps and deblending steps.



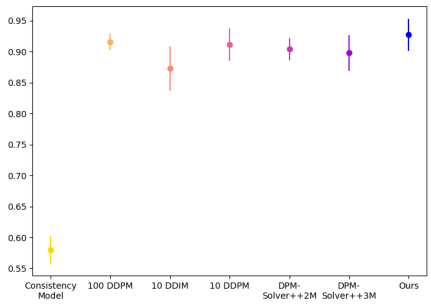
(a) Push-T Rewards

(b) Tool-Hang Reward

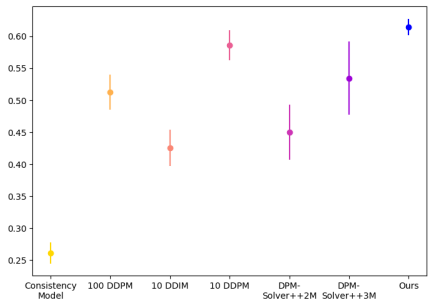
(c) Push-T MSE

(d) Tool-Hang MSE

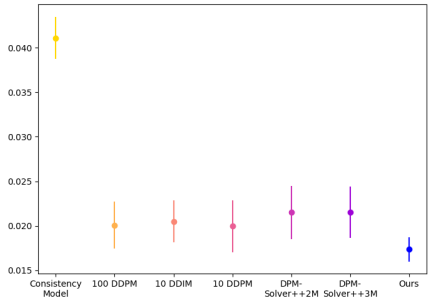
Figure 6: **Results for robot control tasks.** All algorithms except 100 DDPM take a total of 10 steps. ASPeeD is superior to DPM-Solver++ algorithm both in terms of reward and in MSE, for both Tool-Hang and Push-T tasks.



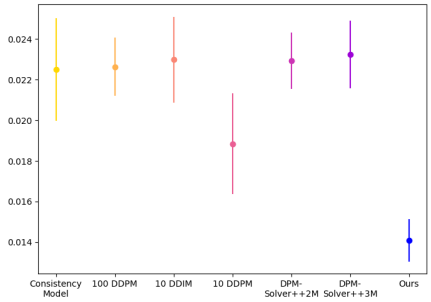
(a) Push-T Rewards



(b) Tool-Hang Reward

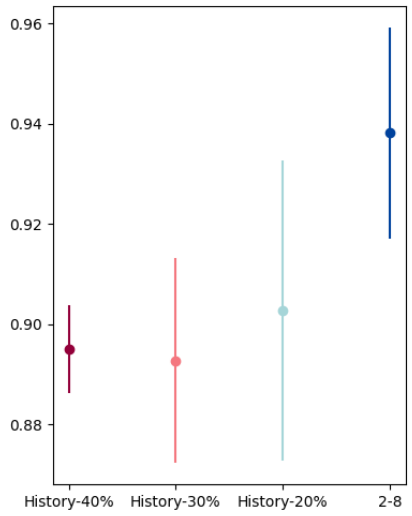


(c) Push-T MSE

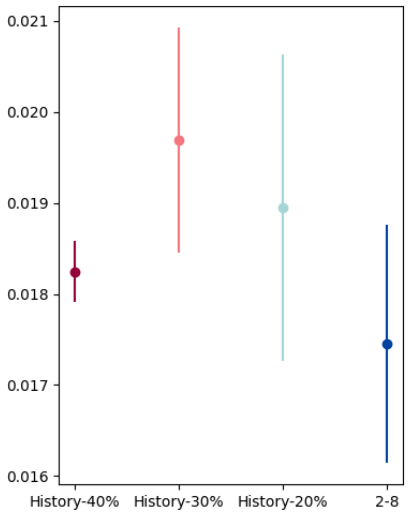


(d) Tool-Hang MSE

Figure 7: **Results for robot control tasks.** Same as Fig. 6 with the addition of a 10-step Consistency Model baseline. Our approach is superior to all other baselines.



(a) Push-T task reward



(b) Push-T task MSE

Figure 8: **Robot control history-based initialization ablations.** Push-T task reward and MSE comparison with different initialization techniques: history-based initialization with 20%, 30% and 40% perturbation and 10 deblending steps, and $DDIM_2$ initialization with 2 DDIM steps and 8 deblending steps (10 total). In this task $DDIM_n$ initialization is superior to history-based initialization.

2021; Song et al., 2020; Ramesh et al., 2022). Of particular interest to our work, diffusion models have been applied to a variety of sequence prediction tasks. In particular, they have been used to varying degrees of success for video prediction (Höppe et al., 2022; Voleti et al., 2022; Yin et al., 2023; Ho et al., 2022; Harvey et al., 2022; Qiu et al., 2019; Yang et al., 2022a,b), as well as for decision making and prediction of robot trajectories (Janner et al., 2022; Chi et al., 2023; Ajay et al., 2023).

One detriment to the usability of diffusion models is inference time, caused by the inherent sequential nature of the denoising process (Song et al., 2020). Many approaches have attempted to alleviate this issue; most notably, DDIM (Song et al., 2020) generalizes the denoising algorithm of DDPM (Ho et al., 2020) to a non-Markovian diffusion process. Both Lu et al. (2022a,b); Zhang & Chen (2022) take advantage of the semi-linear property of the diffusion ODE to use more accurate ODE solvers; Song et al. (2023) recognize the significance of a consistent prediction along a trajectory, while Karras et al. (2022) explore design choices in diffusion algorithms. All these methods trade sample quality for sampling speed. Salimans & Ho (2021) develop a method to distill trained diffusion models and reduce the number of steps required to generate new samples.

Other work aims to perform denoising diffusion starting with non-Gaussian noise, in some cases even taking deterministic steps to denoise non-stochastic transformations. Bansal et al. (2022) propose Cold Diffusion, which aims to learn the reverse process of non-Gaussian image transformations such as blurring or downsampling. Delbracio & Milanfar (2023) take a similar iterative approach to image restoration. Lyu et al. (2022) propose ES-DDPM, which uses samples from a pre-trained model such as a GAN or VAE as a starting point for the denoising process. Closely related to our work are Iterative α-(de)Blending (IADB, Heitz et al. 2023) and Rectified Flow (Liu et al., 2022). Both provide a recipe for blending between two arbitrary distributions. While able to improve on inference time by initializing the diffusion process from distributions other than Gaussian noise, the above approaches center almost exclusively on image generation. In this work we focus on sequence prediction, utilizing the inherent properties and available information in sequences to obtain better initial approximations for the denoising process. Lee et al. (2021) considered an audio domain, and proposed to initialize a non-deterministic diffusion process from a learned Gaussian source distribution. In our work we consider video prediction and robotic control, and explore *non-Gaussian* source distributions, using deterministic diffusion. In parallel, previous studies such as Denton & Fergus (2018) have explored prior selection for video prediction; however, Denton & Fergus (2018) uses variational autoencoders (VAEs), while we focus on the more performant diffusion models.

References

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, Tommi S. Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision making? In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=sP1fo2K9DFG`.

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise, 2022.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion, 2023.

Mauricio Delbracio and Peyman Milanfar. Inversion by direct iteration: An alternative to denoising diffusion for image restoration, 2023.

Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International conference on machine learning*, pp. 1174–1183. PMLR, 2018.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos, 2022.

Eric Heitz, Laurent Belcour, and Thomas Chambon. Iterative α -(de)blending: a minimalist deterministic diffusion model. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings*. ACM, jul 2023. doi: 10.1145/3588432.3591540. URL `https://doi.org/10.1145%2F3588432.3591540`.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022.

Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling, 2022.

Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning*, pp. 9902–9915. PMLR, 2022.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. Priorgrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. In *International Conference on Learning Representations*, 2021.

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv e-prints*, pp. arXiv–2211, 2022a.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, 2022b.

Zhaoyang Lyu, Xudong Xu, Ceyuan Yang, Dahua Lin, and Bo Dai. Accelerating diffusion models via early stop of the diffusion process. *arXiv preprint arXiv:2205.12524*, 2022.

Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xinmei Tian, and Tao Mei. Learning spatio-temporal representation with local and global diffusion, 2019.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2021.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation, 2022.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796*, 2022a.

Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022b.

Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Gong Ming, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. Nuwa-xl: Diffusion over diffusion for extremely long video generation, 2023.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.