DICE: Discrete Inversion Enabling Controllable Editing for Masked Generative Models

Xiaoxiao He[1], Ligong Han[1,2,3†], Quan Dao[1], Song Wen[1], Minhao Bai[1], Di Liu[1], Han Zhang[4], Felix Juefei-Xu[5], Chaowei Tan[1], Bo Liu[6], Martin Renqiang Min[7], Kang Li[1], Faez Ahmed[8], Akash Srivastava[2,3], Hongdong Li[9], Junzhou Huang[10], & Dimitris N. Metaxas[1]
[1]Rutgers [2]MIT-IBM [3] Red Hat AI [4]DeepMind [5]NYU [6]Walmart [7]NEC [8]MIT [9]ANU [10]UTA
† Project Lead & Corresponding Author Project Website: [Link]

Abstract

Recent advances in discrete diffusion models have demonstrated impressive performance in image generation but remain limited in controlled content editing. We propose `DICE` *(Discrete Inversion for Controllable Editing), the first framework to introduce precise inversion for discrete diffusion models, encompassing masked generative and multinomial diffusion variants. Our approach innovatively records Gumbel noise sequences in logit space during reverse diffusion, enabling accurate reconstruction and controlled editing without predefined masks or attention manipulation. Through extensive experiments with image models like Paella and VQ-Diffusion, we validate* `DICE`*'s ability to preserve high fidelity to original data while substantially enhancing editing flexibility, establishing new avenues for fine-grained manipulation in discrete domains.*

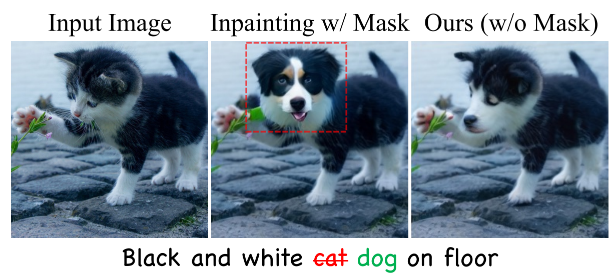

Figure 1. Illustration of the limitation of masked inpainting method. Here, we want to change the cat to a dog. Inpainting with masked generation inadvertently modifies the orientation of the head, resulting in a less favourable result. With our discrete inversion method, we are able to edit the image while preserving other properties of the object being edited. This is achieved by injecting the information from the input image into the logit space. Dotted red box indicates the mask.

1. Introduction

Diffusion models have achieved remarkable success in generating high-fidelity images and videos [8, 18, 33, 34, 36, 38]. These models iteratively denoise samples from noise distributions, reversing a gradual corruption process. Diffusion models are broadly categorized into continuous and discrete types, each tailored to specific data modalities.

Continuous diffusion models utilize stochastic differential equations (SDEs) or ordinary differential equations (ODEs) to handle continuous data [40, 41], with advances such as flow matching [6, 24] enhancing their efficiency. Recent work on visual prompting also improves adaptation efficiency in vision models [22]. Applications include image editing [13, 29], medical imaging [15], and inverse problems [5, 42]. Critical for controlled manipulation is *inversion*, which recovers latent representations through deterministic [40] or stochastic inversion [8, 47].

Discrete diffusion models target inherently discrete data like text or image tokens [1, 11, 19]. Prominent examples include multinomial diffusion [19] and masked generative models like MaskGIT [3]. However, controlled content editing remains challenging. Masked generative models use masked inpainting for editing but lack fine-grained control, as regions cannot inject information into regenerated areas (Figure 1). Moreover, ODE-based inversion techniques do not directly apply to discrete models due to fundamental differences in data representation and diffusion processes.

To address these limitations, we propose `DICE` (Discrete Inversion for Controllable Editing), the first inversion algorithm for discrete diffusion models to our knowledge. Our method extends stochastic inversion to discrete diffusion models, including multinomial diffusion and masked generative models, by recording noise sequences during reverse diffusion. Specifically, we construct artificial trajectories with low latent-state correlation, fit reverse sampling steps, and record residuals between targets and predictions. These residuals *imprint* original data information, enabling con-

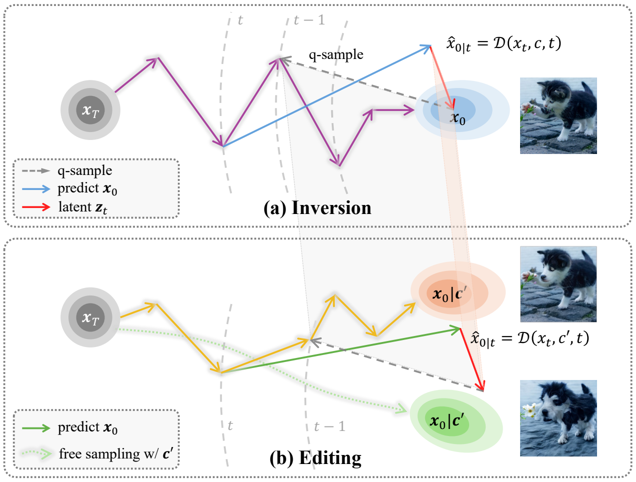

Figure 2. Inversion and editing process for masked generative modeling (MGM) as in Algorithm 1.

trolled editing by reinjecting residuals during inference.

DICE achieves accurate reconstruction and controlled editing without predefined masks or attention manipulation, providing flexibility for fine-grained discrete data editing. We validate our method extensively across image and text modalities, demonstrating effectiveness with models like VQ-Diffusion [11] and Paella [37]. Additionally, we introduce a new text-editing dataset to facilitate future research. Our contributions are: (1) Introducing DICE, enabling precise inversion and controlled editing for discrete diffusion models through stochastic inversion with noise residual recording. (2) Demonstrating DICE's effectiveness and versatility across image and text generative models through extensive experiments.

2. Related Work

Discrete diffusion. Diffusion models in discrete spaces were initially explored in [39], and further formalized using discrete-time Markov chains by Argmax Flows [19] and D3PM [1]. These models reverse the noising process via variational training. VQ-GAN [9, 11] enables token-based image generation, paving the way for non-autoregressive models like MaskGIT [3], Muse [4], and MMVID [12]. Recent methods adapt score matching to discrete spaces, such as ratio matching [26, 30] and discrete flow matching [10]. In NLP, masked language models like BERT [7] and RoBERTa [25] have also been interpreted as discrete diffusion processes [44].

Diffusion inversion. Inversion aims to recover the latent code that reconstructs input data. Deterministic methods like DDIM [40] and flow matching [24] use neural ODEs, while stochastic approaches like DDPM Inversion [20] and CycleDiffusion [47] trace noise in SDEs. Our method generalizes DDPM Inversion to discrete diffusion models, bridging the gap between continuous and discrete domains.

Inversion-based image editing. DDIM inversion underpins many editing techniques, often combined with attention guidance [16, 17, 27, 43]. DDPM inversion methods [20] offer simpler interfaces and integrate with semantic guidance like SEGA and LEDITS++. Null-text Inversion [32] enhances fidelity through test-time embedding optimization, while Negative-prompt Inversion [14, 31] offers efficient, closed-form solutions to improve editing speed.

3. Methods

3.1. Preliminaries

Masked generative modeling. Masked generative modeling is widely used in representation learning for both natural language processing and computer vision. It works by masking parts of the input and training the model to reconstruct the missing data. In models like BERT [7] and RoBERTa [25], masked tokens (`[MASK]`) are predicted based on the surrounding context, excelling in text completion and embedding representation learning. For image generation, Paella [37] adapts this approach for text-conditional image generation by renoising tokens instead of masking. The inference process in masked generative models typically involves iterative renoise/remask and repredict steps.

Multinomial Diffusion. Denoting $\boldsymbol{x}_0 \in \{1,\dots,K\}^D$ as a data point of dimension D. We use $\boldsymbol{v}(x_t^{(i)})$ to denote the one hot column vector representation of the i-th entry of $\boldsymbol{x}_t$. To simplify notation, in the following we drop index i and any function that operates on vector $\boldsymbol{x}_t$ is populated along its dimension. Diffusion model defines a markov chain $q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \Pi_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ that gradually add noise to the data $\boldsymbol{x}_0$ for T times so that $\boldsymbol{x}_T$ contains little to no information. Discrete diffusion model [1, 11, 19] proposed an alternative likelihood-based model for categorical data, and defines the forward process following:

$$q(x_t|x_{t-1}) = \mathrm{Cat}\left(\boldsymbol{v}(x_t); \boldsymbol{\pi} = \boldsymbol{Q}_t \boldsymbol{v}(x_{t-1})\right). \tag{1}$$

where $\boldsymbol{Q}_t$ is the transition matrix between adjacent states following mask-and-replace strategy. The posterior distribution given x_0 has a closed-form solution,

$$q\left(x_{t-1}|x_t, x_0\right) = \frac{(\boldsymbol{Q}_t^\top \boldsymbol{v}(x_t)) \odot (\overline{\boldsymbol{Q}}_{t-1}\boldsymbol{v}(x_0))}{\boldsymbol{v}(x_t)^\top \overline{\boldsymbol{Q}}_t \boldsymbol{v}(x_0)}. \tag{2}$$

where $\overline{\boldsymbol{Q}}_t = \boldsymbol{Q}_t \cdots \boldsymbol{Q}_1$ is the cumulative transition matrix. The details of $\boldsymbol{Q}_t$ and $\overline{\boldsymbol{Q}}_t$ are given in the supplementary materials. The inference process is as below:

$$\boldsymbol{\pi}_\theta(x_t, t) = p_\theta\left(x_{t-1}|x_t\right) = \sum_{\tilde{x}_0=1}^{K} q\left(x_{t-1}|x_t, \tilde{x}_0\right) p_\theta\left(\tilde{x}_0|x_t\right), \tag{3}$$

with $p_\theta(\tilde{x}_0|x_t)$ is parameterized by a neural network. We gradually denoise from x_T to x_0 using 3. For numerical stability, the implementation uses log space instead of probability space. Masked generative models can be viewed as a special case of multinomial diffusion models with an additional *absorbing* state (or the `[MASK]` state). Its training objective can be viewed as a reweighted ELBO [2].

3.2. Discrete Inversion for Controllable Editing

Non ODE-based inversion. ODE-based generative models, such as DDIM and flow matching, define an ODE trajectory. Due to the deterministic nature of ODEs, inversion can be achieved by solving the ODE using the Euler method in forward direction, ensuring reconstruction based on the inherent properties of the ODE. In contrast, another line of research focuses on SDE-based models, such as CycleDiffusion [47] and DDPM Inversion [20]. Broadly speaking, these approaches ensure reconstruction by recording the noises or residuals that are required to reproduce the stochastic trajectory. CycleDiffusion records the Gaussian noise $\boldsymbol{z}_t$ during sampling from posterior $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0 = \boldsymbol{x}_0)$ and injects information of the input signal by feeding the true $\boldsymbol{x}_0$. DDPM Inversion, on the other hand, incorporates information into $\boldsymbol{z}_t$ by fitting the reverse process into an artificial stochastic trajectory obtained by independent `q-sample`. For both CycleDiffusion and DDPM Inversion, the key idea is to utilize the Gaussian reparameterization trick, $x = \mu + \sigma z \Leftrightarrow x \sim \mathcal{N}(x; \mu, \sigma^2)$, and keeping track of the "noise" that could have generated the sample from mean. For discrete diffusion models, we utilize the Gumbel-Max trick [21, 28], $x = \arg\max(\log(\boldsymbol{\pi}) + \boldsymbol{g}) \Leftrightarrow x \sim \mathrm{Cat}(x; \boldsymbol{\pi})$. Figure 2 provides an intuition of the proposed method.

Method	PSNR ↑	LPIPS$_{\times 10^3}$ ↓	MSE$_{\times 10^4}$ ↓	SSIM$_{\times 10^2}$ ↑
Inpainting	10.50	565.11	1002.09	30.13
Ours	**30.91**	**39.81**	**11.07**	**90.22**
Ours†	Inf	0.07	0.01	99.99

Table 1. **Inversion Reconstruction performance.** † The metric is calculated between the original image and its inverted counterpart. Due to the encoding and decoding steps in the VQ-VAE/GAN process, some inaccuracies are introduced by the quantization. The PSNR is `Inf` due to the reconstruction of our method yielding the same VQ-VAE/GAN latents. Base model is Paella [37].

Inverting masked generative models. In masked generative modeling, the stochastic trajectory $\boldsymbol{x}_t$ is constructed according to the specific inference algorithm of the model in use. For example, in Paella [37], the masking is *inclusive*, meaning that as the time step t increases, the set of masked tokens grows. In contrast, the Unleashing Transformer [2] employs *random* masking at each step, where masks are generated independently using the `q-sample`

Algorithm 1 Discrete Inversion for Masked Generative Modeling

Inversion:
1: $\boldsymbol{y}_0 \leftarrow \mathcal{D}(\boldsymbol{x}_0, \boldsymbol{c}, t = 0)$
2: Sample noise token map $\boldsymbol{n}$
3: **for** t from 1 to T **do**
4: $\quad \boldsymbol{m}_t \leftarrow \mathrm{GenerateMask}(t)$ ▷ Sampling masks according to inference algorithm
5: $\quad \boldsymbol{x}_t \leftarrow \boldsymbol{x}_0 \odot (\boldsymbol{1} - \boldsymbol{m}_t) + \boldsymbol{n} \odot \boldsymbol{m}_t$
6: $\quad \hat{\boldsymbol{y}}_{0|t} \leftarrow \mathcal{D}_\theta(\boldsymbol{x}_t, \boldsymbol{c}, t = t)$
7: $\quad \boldsymbol{z}_t \leftarrow \boldsymbol{y}_0 - \hat{\boldsymbol{y}}_{0|t}$ ▷ Eq 4
8: **end for**
Editing/Sampling:
9: **for** t from τ to 1 **do**
10: $\quad \hat{\boldsymbol{y}}_{0|t} \leftarrow \mathcal{D}_\theta(\boldsymbol{x}_t, \boldsymbol{c}', t = t)$
11: $\quad \boldsymbol{g} \sim \mathrm{Gumbel}(\boldsymbol{0}, \boldsymbol{I})$
12: $\quad \tilde{\boldsymbol{y}}_0 \leftarrow \hat{\boldsymbol{y}}_{0|t} + \lambda_1 \cdot \boldsymbol{z}_t + \lambda_2 \cdot \boldsymbol{g}$
13: $\quad \tilde{\boldsymbol{x}}_0 \leftarrow \arg\max \tilde{\boldsymbol{y}}_0$
14: $\quad \boldsymbol{x}_{t-1} \leftarrow \tilde{\boldsymbol{x}}_0 \odot (\boldsymbol{1} - \boldsymbol{m}_{t-1}) + \boldsymbol{n} \odot \boldsymbol{m}_{t-1}$ ▷ Re-noise
15: **end for**
16: Return $\boldsymbol{x}_0$.

function. Without loss of generality, we define a denoiser function $\mathcal{D}_\theta$ (parameterized by θ). This denoiser outputs the *logits* of the predicted unmasked data given the noisy tokens $\boldsymbol{x}_t$. Unlike DDPM or multinomial diffusion, where x_{t-1} is *not* sampled from a posterior given x_t, the inference of masked modeling takes a different approach. In masked modeling, x_t is obtained from sampled $\hat{x}_{0|t}$ by re-noising. Since the categorical sampling happens at sampling from the denoiser's prediction, we therefore define an corresponding latent sequence:

$$\begin{aligned} \hat{\boldsymbol{y}}_{0|t} &= \log(p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_t)) = \mathcal{D}_\theta(\boldsymbol{x}_t, t) \\ \boldsymbol{z}_t &:= \boldsymbol{y}_0 - \hat{\boldsymbol{y}}_{0|t}. \end{aligned} \tag{4}$$

With our proposed latent space, accurate reconstruction is guaranteed. However, for editing tasks, this level of precision may not be ideal if the latent variable $\boldsymbol{z}_t$ dominates the generation process. The detailed algorithm is given in Algorithm 1.

To provide more flexibility, we introduce the hyperparameters τ, λ_1, and λ_2, which allow for finer control over the editing process. Specifically, τ represents the starting (and largest) timestep at which the editing process begins, while λ_1 controls the amount of information injected from the original input, and λ_2 governs the introduction of random noise (Algorithm 1 line 12).

4. Experiments

We evaluate the effectiveness of `DICE` on image diffusion models. Experiments show that our method preserves identity while enabling controlled editing. See Supplementary Materials for implementation details.

Image diffusion models and dataset. We evaluate on absorbing-state discrete models [1], including the masked generative model Paella and multinomial diffusion model VQ-Diffusion. For benchmarking, we use the Prompt-based Image Editing Benchmark (PIE-Bench) [23], which assesses text-to-image editing across 9 scenarios with 700 annotated images.

Reconstruction and editing evaluation. We first assess reconstruction quality by comparing original and reconstructed images using PSNR, LPIPS, MSE, and SSIM. As shown in Table 1, DICE achieves near-perfect reconstruction, far outperforming Inpainting + Paella, which lacks access to the original image structure due to full masking. Our method avoids quantization artifacts seen in VQ-VAE/GAN baselines, highlighting its fidelity and consistency.

We then evaluate editing performance using eight metrics across three aspects: structural similarity [43], background preservation (PSNR, LPIPS [48], MSE, SSIM [45]), and semantic alignment via CLIP [35] similarity [46]. Table 2 shows DICE with Paella achieves the lowest structure distance (11.34), outperforming even continuous diffusion baselines. While Stable Diffusion scores higher on CLIP similarity, our method offers better structural fidelity, achieving a strong trade-off between prompt alignment and preservation of image content.

Using VQ-Diffusion, DICE also shows robust editing performance. As seen in Table 3, our method significantly outperforms DDIM+SD1.4 in preserving unedited regions across all background metrics. These results confirm that original image information is effectively encoded and reinjected during editing. Figure 3 presents visual examples with Paella. Our method consistently modifies real images according to target prompts while maintaining high fidelity to the original content.

5. Conclusion

In this paper, we introduced an inversion algorithm for discrete diffusion models, including multinomial diffusion and masked generative models. By leveraging recorded noise sequences and masking patterns during the reverse diffusion process, DICE enables accurate reconstruction and flexible editing of discrete data without the need for predefined masks or cross-attention manipulation. Our experiments across multiple models and modalities demonstrate its effectiveness in preserving data fidelity while enhancing editing capabilities. We believe that DICE enhances the capabilities of discrete generative models, offering new opportunities for fine-grained content manipulation.

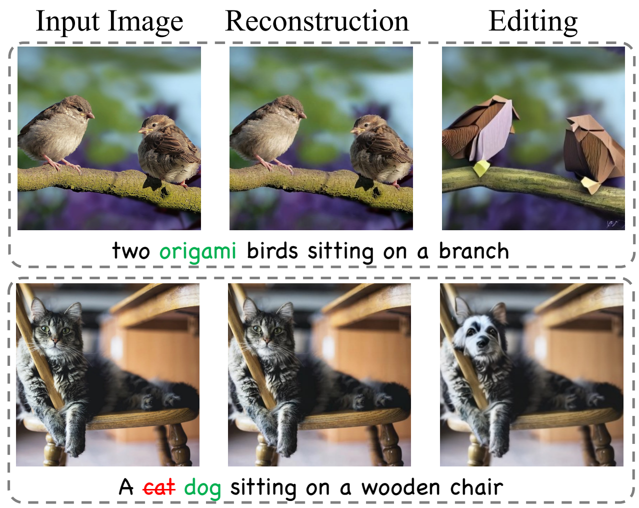

Figure 3. **Visualization of editing results.** Editing results for our method using Paella, along with their corresponding prompts.

	Method		Structure	CLIP Similarity	
	Inversion+Model	Editing	Distance$_{\times 10^3}$ ↓	Whole ↑	Edited ↑
Continuous	DDIM+SD1.4	P2P	69.43*	25.01*	22.44*
	Null-Text + SD1.4	P2P	13.44*	24.75*	21.86*
	Negative-Prompt + SD1.4	P2P	16.17*	24.61*	21.87*
	DDPM-Inversion + SD1.4	Prompt	22.12	**26.22**	<u>23.02</u>
Discrete	Inpainting + Paella	Prompt	91.10	<u>25.36</u>	**23.42**
	Ours + Paella	Prompt	**11.34**	23.79	21.23
	Ours + VQ-Diffusion†	Prompt	<u>12.70</u>	23.85	21.02

Table 2. **Quantitative results on image editing performance.** Comparison of DICE with the masked inpainting with the discrete diffusion models as well as continuous ones (Stable Diffusion v1.4) using DDIM inversion. "P2P" refers to Prompt-to-Prompt [16], and "Prompt" denotes editing performed solely through forward edit prompts. Entries marked with an asterisk (*) are cited from [23]. †: For VQ-Diffusion, the images are downsampled to 256 × 256. Please note that due to differences in base models and editing algorithms, the metrics across methods are not directly comparable. However, our method significantly outperforms both inpainting and strong baselines (e.g., Null-Text Inversion + SD1.4) in terms of structural preservation. As expected, inpainting achieves a high CLIP score since it directly generates image patches based on the target prompt.

Method		Background Preservation			
Inversion+Model	Editing	PSNR ↑	LPIPS$_{\times 10^3}$ ↓	MSE$_{\times 10^4}$ ↓	SSIM$_{\times 10^2}$ ↑
DDIM+SD1.4	P2P	17.87	208.80	219.88	71.14
Ours+Paella	Prompt	**27.29**	**52.90**	**43.76**	**89.79**

Table 3. **Background Preservation.** Quantitative comparison of background preservation between our proposed method and DDIM+SD 1.4, achieved by masking the edited region and calculating image similarity with the unedited masked image. The inpainting is served as upper bound since only the masked region are edited and background are not modified.

References

[1] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34: 17981–17993, 2021. 1, 2, 4

[2] Sam Bond-Taylor, Peter Hessey, Hiroshi Sasaki, Toby P Breckon, and Chris G Willcocks. Unleashing transformers: Parallel token prediction with discrete absorbing diffusion for fast high-resolution image generation from vector-quantized codes. In *European Conference on Computer Vision*, pages 170–188. Springer, 2022. 3

[3] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11315–11325, 2022. 1, 2

[4] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 2

[5] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 1

[6] Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv preprint arXiv:2307.08698*, 2023. 1

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[9] Patrick Esser, Robin Rombach, Andreas Blattmann, and Bjorn Ommer. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in neural information processing systems*, 34:3518–3532, 2021. 2

[10] Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *arXiv preprint arXiv:2407.15595*, 2024. 2

[11] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022. 1, 2

[12] Ligong Han, Jian Ren, Hsin-Ying Lee, Francesco Barbieri, Kyle Olszewski, Shervin Minaee, Dimitris Metaxas, and Sergey Tulyakov. Show me what and tell me how: Video synthesis via multimodal conditioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3615–3625, 2022. 2

[13] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*, 2023. 1

[14] Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, et al. Proxedit: Improving tuning-free real image editing with proximal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4291–4301, 2024. 2

[15] Xiaoxiao He, Chaowei Tan, Ligong Han, Bo Liu, Leon Axel, Kang Li, and Dimitris N Metaxas. Dmcvr: Morphology-guided diffusion model for 3d cardiac volume reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 132–142. Springer, 2023. 1

[16] 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, 4

[17] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 2

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[19] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. *Advances in Neural Information Processing Systems*, 34:12454–12465, 2021. 1, 2

[20] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. An edit friendly ddpm noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12469–12478, 2024. 2, 3

[21] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 3

[22] Can Jin, Ying Li, Mingyu Zhao, Shiyu Zhao, Zhenting Wang, Xiaoxiao He, Ligong Han, Tong Che, and Dim-

itris N. Metaxas. Lor-VP: Low-rank visual prompting for efficient vision model adaptation. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[23] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023. 4

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 1, 2

[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[26] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion language modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023. 2

[27] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023. 2

[28] Chris J Maddison, Daniel Tarlow, and Tom Minka. A* sampling. *Advances in neural information processing systems*, 27, 2014. 3

[29] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 1

[30] Chenlin Meng, Kristy Choi, Jiaming Song, and Stefano Ermon. Concrete score matching: Generalized score matching for discrete data. *Advances in Neural Information Processing Systems*, 35:34532–34545, 2022. 2

[31] Daiki Miyake, Akihiro Iohara, Yu Saito, and Toshiyuki Tanaka. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. *arXiv preprint arXiv:2305.16807*, 2023. 2

[32] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 2

[33] OpenAI. Sora: Video generation model, 2024. Accessed: 2024-10-09. 1

[34] Hao Phung, Quan Dao, Trung Dao, Hoang Phan, Dimitris Metaxas, and Anh Tran. Dimsum: Diffusion mamba–a scalable and unified spatial-frequency method for image generation. *arXiv preprint arXiv:2411.04168*, 2024. 1

[35] 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. 4

[36] 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

[37] Dominic Rampas, Pablo Pernias, and Marc Aubreville. A novel sampling scheme for text-and image-conditional image synthesis in quantized latent spaces. *arXiv preprint arXiv:2211.07292*, 2022. 2, 3

[38] 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)*, pages 10684–10695, 2022. 1

[39] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 1, 2

[41] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1

[42] Anastasis Stathopoulos, Ligong Han, and Dimitris Metaxas. Score-guided diffusion for 3d human recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 906–915, 2024. 1

[43] 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, 4

[44] Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*, 2019. 2

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[46] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan.

Godiva: Generating open-domain videos from natural descriptions. *arXiv preprint arXiv:2104.14806*, 2021. 4

[47] Chen Henry Wu and Fernando De la Torre. Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. *arXiv preprint arXiv:2210.05559*, 2022. 1, 2, 3

[48] 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. 4