KV-Edit: Training-Free Image Editing for Precise Background Preservation

Supplementary Material

In this supplementary material, we provide more details and findings. In Appendix A, we present additional experimental results and implementation details of our proposed KV-Edit. Appendix B provides further discussion and data regarding our inversion-free methodology. Appendix C details the design and execution of our user study. In Appendix D, we clarify the differences between our method and MasaCtrl [9]. Finally, in Appendix E, we discuss potential future directions and current limitations of our work.

A. Implementation and More Experiments

Implementation Details. Our code is built on Flux [1], with modifications to both double block and single block to incorporate KV cache through additional function parameters. Input masks are first downsampled using bilinear interpolation, then transformed from single-channel to 64-channel representations following the VAE in Flux [1]. In the feature space, the smallest pixel unit is 16 dimensions rather than the entire 64-dimensional token. Therefore, in addition to KV cache, we preserve the intermediate image features at each timestep to ensure fine-grained editing capabilities. In our experiment, inversion and denoising can be performed independently, allowing a single image to be inverted just once and then edited multiple times with different conditions, further enhancing the practicality of this workflow.

Experimental Results. Due to space constraints in the main paper, we only present results on the PIE-Bench [23]. Here, we provide additional examples demonstrating the effectiveness of our approach. To further showcase the flexibility of our method, Fig. A and Fig. B present various editing target applied to the same source image, without explicitly labeling the input masks because each case corresponds to a different mask. Fig. D illustrates the impact of steps and reinitialization strategy on the color changing tasks and inpainting tasks.

When changing colors, as the number of skip-steps decreases and reinitialization strategy is applied, the color information in the tokens is progressively disrupted, ultimately achieving successful results. In our experiments, the optimal number of steps to skip depends on image resolution and content, which can be adjusted based on specific needs and feedback. Unlike previous training-free methods, our approach even can be applied to inpainting tasks after employing reinitialization strategy, as demonstrated in the third row of Fig. D. The originally removed regions in inpainting tasks can be considered as black objects, thus requiring reinitialization strategy to eliminate pure black in-

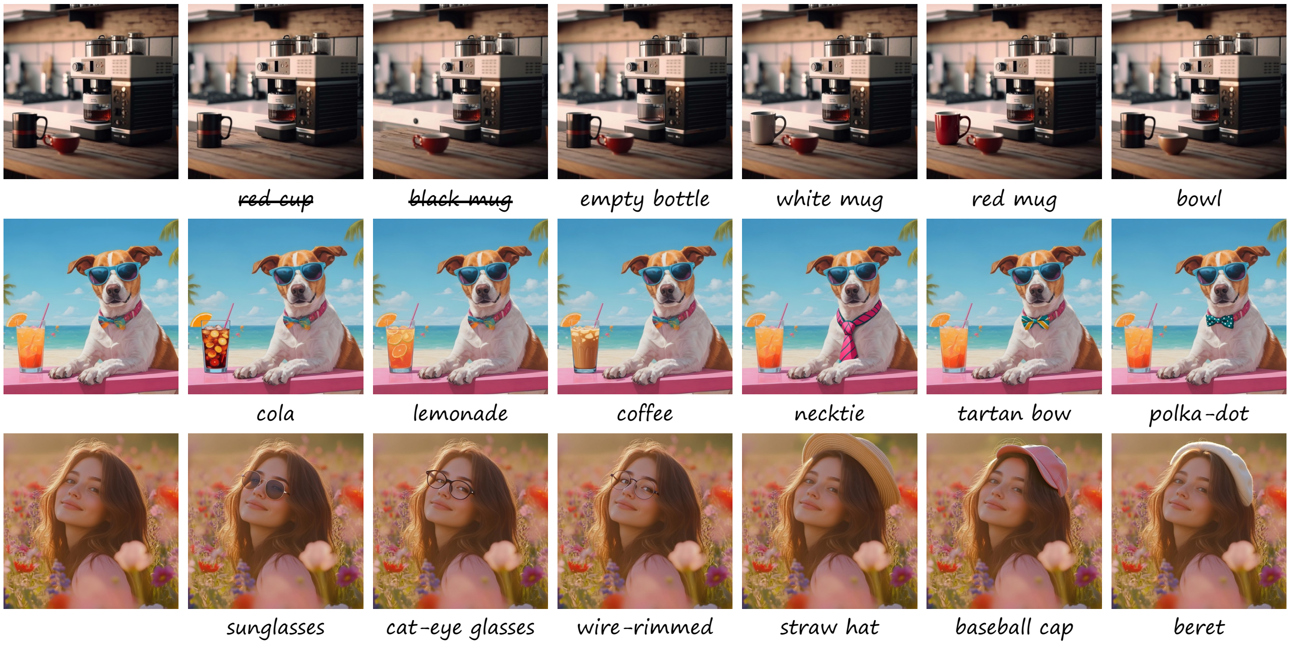


Figure A. **Additional editing results of KV-Edit**. Our method demonstrates robust performance with strict background preservation and high image quality. Users can achieve creative designs by simply adjusting text prompts and masks according to their needs.

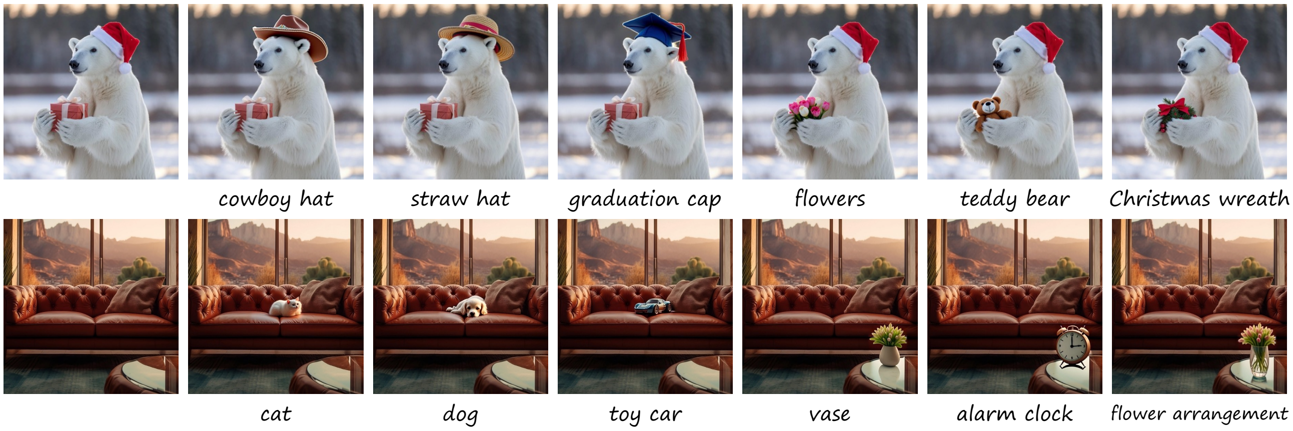


Figure B. **Additional editing results of KV-Edit**. Our method demonstrates robust performance with strict background preservation and high image quality. Users can achieve creative designs by simply adjusting text prompts and masks according to their needs.

formation and generate meaningful content. We plan to further extend our method to inpainting tasks in future work, as there are currently very few training-free methods available for this application.

Attention Scale When dealing with large masks (e.g., background changing tasks), our original method may produce discontinuous images including conflicting content, as illustrated in Fig. C. Stable-Flow [4] demonstrated that during image generation with DiT [48], image tokens primarily attend to their local neighborhood rather than globally across most layers and timesteps.

Consequently, although our approach treats the background as a condition to guide new content generation, large masks can introduce generation bias which ignore existing content and generate another objects. Based on this analysis, we propose a potential solution as shown in Fig. C. We directly increase the attention weights from masked regions to unmasked regions in the attention map (produced by query-key multiplication), effectively mitigating the bias impact. This attention scale mechanism enhances content coherence by strengthening the influence of preserved background on new content.

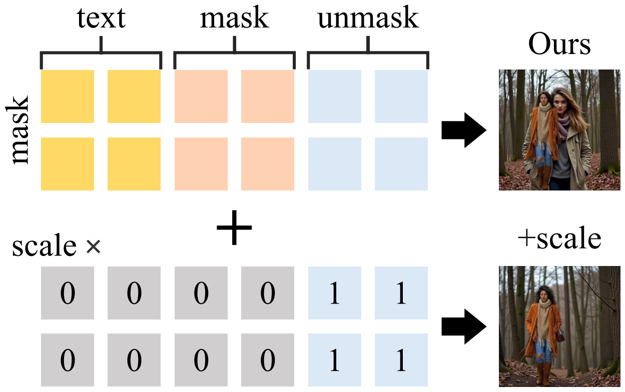


Figure C. **Implementation of attention scale**. The scale can be adjusted to achieve optimal results.

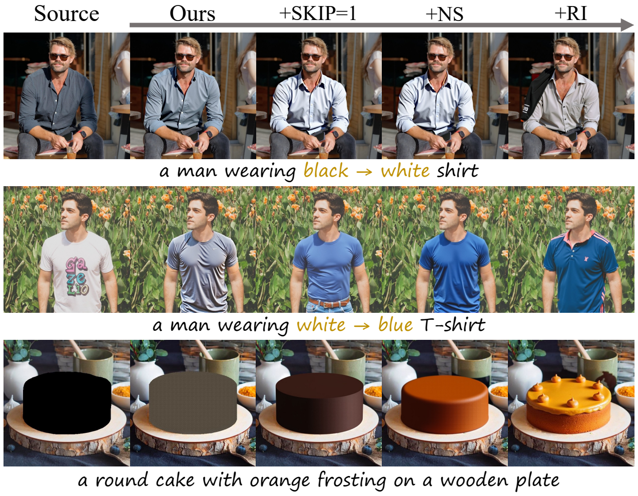

Source
Ours
+SKIP=1
+NS
+RI
a man wearing black → white shirt
a man wearing white → blue T-shirt
a round cake with orange frosting on a wooden plate

Figure D. **Additional ablation studies on two tasks**. The first and second rows demonstrate the impact of timesteps and reinitialization strategy (**RI**) on color changing. The third row demonstrates the impact of timesteps and **RI** on the inpainting tasks.

B. More Discussions on Inversion-Free

We implement inversion-free editing on Flux [1] based on the code provided by FlowEdit [25]. As noted in FlowEdit [25], adding random noise at each editing step may introduce artifacts, a phenomenon we also demonstrate in the main paper. In this section, we primarily explore the impact of inversion-free methods on memory consumption.

Algorithm A demonstrates the implementation of inversion-free KV-Edit, where "KV-inversion" and "KV-denoising" refer to single-step noise prediction with KV cache. KV cache is saved during a one-time inversion process and immediately utilized in the denoising process. The final vector can be directly added to the original image without first inversing it to noise. This strategy ensures that the space complexity of KV cache remains $O(1)$ along the time dimension. Moreover, resolution has a more significant im-

timesteps	512 × 512		768 × 768	
	Ours	+Inf.	Ours	+Inf.
24 steps	16.2G	**1.9G**	65.8G	3.5G
28 steps	19.4G	**1.9G**	75.6G	3.5G
32 steps	22.1G	**1.9G**	86.5G	3.5G

Table A. **Memory usage at different resolutions and timesteps**. Our approach has a space complexity of $O(n)$ along the time dimension, while inversion-free methods achieve $O(1)$.

Resolution	TMACs ↓			
	Vanilla	DiffEdit	RF Edit	Ours
512 × 512	555.5	1018.7	1111.8	**425.6**
768 × 768	1018.7	1018.7	2037.4	**722.6**

Table B. **Computational efficiency at different resolutions tested on RTX3090.** The baseline is a vanilla ReFlow model utilizing 28 steps for both inversion and denoising.

Algorithm A Simplified Inf. version KV-Edit

```
**Input:** $t_i$, real image $x_0^{src}$, foreground $z_{t_i}^{fg}$,foreground region $mask$, KV cache $C$
**Output:** Prediction vector $\boldsymbol{v}_{\theta t_i}^{fg}$
$N_{t_i} \sim \mathcal{N}(0,1)$
$x_{t_i}^{src} = (1-t_i)x_{t_0}^{src} + t_i N_{t_i}$
$\boldsymbol{v}_{\theta t_i}^{src}, C = \text{KV-Inverison}(x_{t_i}^{src}, t_i, C)$
$\tilde{z}_{t_i}^{fg} = z_{t_i}^{fg} + mask \cdot (x_{t_i}^{src} - x_0^{src})$
$\tilde{\boldsymbol{v}}_{\theta t_i}^{fg}, C = \text{KV-Denosing}(\tilde{z}_{t_i}^{fg}, t_i, C)$
**Return** $\boldsymbol{v}_{\theta t_i}^{fg} = \tilde{\boldsymbol{v}}_{\theta t_i}^{fg} - \boldsymbol{v}_{\theta t_i}^{src}$
```

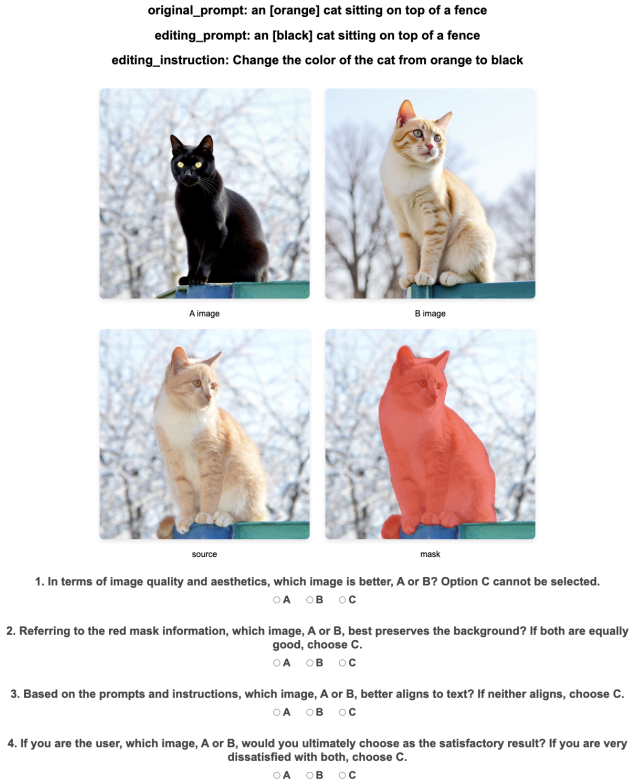
pact on memory consumption as the number of image tokens grows at a rate of $O(n^2)$.

We conducted experiments across various resolutions and time steps, reporting memory usage in Tab. A. When processing high-resolution images and more timesteps, personal computers struggle to accommodate the memory requirements. Nevertheless, we still recommend the inversion-based KV-Edit approach for several reasons:

1. Current inversion-free methods occasionally introduce artifacts.
2. Inversion-based KV-Edit enables multiple editing attempts after a single inversion, significantly improving usability and workflow efficiency.
3. Large generative models inherently require substantial GPU memory, which presents another challenge for personal computers. Therefore, we position inversion-based KV-Edit as a server-side technology.

C. User Study Details

We conduct our user study in a questionnaire format to collect user preferences for different methods. We observe that in most cases, users struggle to distinguish the background effects of training-based inpainting methods (e.g., FLUX-Fill [1] sometimes increases grayscale tones in images). Therefore, we allowed participants to select "equally good" regarding background quality.



original_prompt: an [orange] cat sitting on top of a fence

editing_prompt: an [black] cat sitting on top of a fence

editing_instruction: Change the color of the cat from orange to black

1. In terms of image quality and aesthetics, which image is better, A or B? Option C cannot be selected.

○A ○B ○C

2. Referring to the red mask information, which image, A or B, best preserves the background? If both are equally good, choose C.

○A ○B ○C

3. Based on the prompts and instructions, which image, A or B, better aligns to text? If neither aligns, choose C.

○A ○B ○C

4. If you are the user, which image, A or B, would you ultimately choose as the satisfactory result? If you are very dissatisfied with both, choose C.

○A ○B ○C

Figure E. **User study**. We provide a sample where participants were presented with the original image, editing prompts, results from two different methods for comparison and four questions from four aspects.

Additionally, PIE-Bench [23] contains several challenging cases where all methods fail to complete the editing tasks satisfactorily. Consequently, we allow users to select "neither is good" for text alignment and overall satisfaction metrics, as illustrated in Fig. E.

We implement a single-blind mechanism where the corresponding method for each question is randomly sampled, ensuring fairness in the comparison. We collect over 2,000 comparison results and calculate our method's win rate after excluding cases where both methods are rated equally.

D. Difference with MasaCtrl

We adopt a similar idea to MasaCtrl [9], separating the foreground and background. However, MasaCtrl overlooks three factors that affect the background content: errors, text, and foreground. The method proposed by MasaCtrl not only fails to reduce errors but also disrupts the attention interaction between the foreground and background, resulting in completely different background content in the final results. In contrast, we analyze these issues and successfully separate the foreground and background, ensuring strict consistency. Regarding the implementation, our core idea is to split the query, instead of replacing key and value, which is more reasonable and better motivated.

E. Limitations and Future Work

In this section, we outline the current challenges faced by our method and potential future improvements. While our approach effectively preserves background content, it struggles to maintain foreground details. As shown in Fig. D, when editing garment colors, clothing appearance features may be lost, such as the style, print or pleats.

Typically, during the generation process, early steps determine the object's outline and color, with specific details and appearance emerging later. In the contrast, during inversion, customized object details are disrupted first and subsequently influenced by new content during denoising. This represents a common challenge in the inversion-denoising paradigm [13, 18, 59].

In future work, we could employ trainable tokens to preserve desired appearance information during inversion and inject it during denoising, still without fine-tuning of the base generative model. Furthermore, our method could be adapted to other modalities, such as video and audio editing, image inpainting tasks. We hope that "KV cache for editing" can be considered an inherent feature of the DiT [48] architecture.

References

[1] Flux. `https://github.com/black-forest-labs/flux/`. 2, 3, 5, 6, 7, 8, 9, 10, 11

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *TOG*, 42(4):1–11, 2023. 2

[4] Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. *arXiv preprint arXiv:2411.14430*, 2024. 2, 5, 10

[5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 3

[6] Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, and Yansong Tang. Self-calibrated clip for training-free open-vocabulary segmentation. *arXiv preprint arXiv:2411.15869*, 2024. 3

[7] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 2, 3

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 3

[9] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, pages 22560–22570, 2023. 2, 5, 6, 7, 9, 11

[10] Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. *arXiv preprint arXiv:2409.01055*, 2024. 8

[11] Zhennan Chen, Yajie Li, Haofan Wang, Zhibo Chen, Zhengkai Jiang, Jun Li, Qian Wang, Jian Yang, and Ying Tai. Region-aware text-to-image generation via hard binding and soft refinement. *arXiv preprint arXiv:2411.06558*, 2024. 2

[12] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *ECCV*, pages 390–408, 2024. 2

[13] Wenkai Dong, Song Xue, Xiaoyue Duan, and Shumin Han. Prompt tuning inversion for text-driven image editing using diffusion models. In *ICCV*, pages 7430–7440, 2023. 2, 3, 5, 12

[14] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[15] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 2

[16] Kunyu Feng, Yue Ma, Bingyuan Wang, Chenyang Qi, Haozhe Chen, Qifeng Chen, and Zeyu Wang. Dit4edit: Diffusion transformer for image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2969–2977, 2025. 8

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 3

[18] 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, 3, 5, 6, 7, 12

[19] Wenke Huang, Jian Liang, Zekun Shi, Didi Zhu, Guancheng Wan, He Li, Bo Du, Dacheng Tao, and Mang Ye. Learn from downstream and be yourself in multimodal large language model fine-tuning. *arXiv preprint arXiv:2411.10928*, 2024. 3

[20] Xiaoke Huang, Jianfeng Wang, Yansong Tang, Zheng Zhang, Han Hu, Jiwen Lu, Lijuan Wang, and Zicheng Liu.

Segment and caption anything. In *CVPR*, pages 13405–13417, 2024. 3

[21] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008. 7

[22] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *ECCV*, pages 150–168, 2024. 2, 3, 7

[23] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Pnp inversion: Boosting diffusion-based editing with 3 lines of code. In *ICLR*, 2024. 2, 6, 7, 8, 9, 11

[24] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pages 6007–6017, 2023. 2, 3

[25] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. *arXiv preprint arXiv:2412.08629*, 2024. 2, 5, 6, 7, 10

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023. 3

[27] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing. *arXiv preprint arXiv:2303.15649*, 2023. 2

[28] Yaowei Li, Yuxuan Bian, Xuan Ju, Zhaoyang Zhang, Ying Shan, and Qiang Xu. Brushedit: All-in-one image inpainting and editing. *arXiv preprint arXiv:2412.10316*, 2024. 2, 3, 6, 7, 8

[29] Haonan Lin, Mengmeng Wang, Jiahao Wang, Wenbin An, Yan Chen, Yong Liu, Feng Tian, Guang Dai, Jingdong Wang, and Qianying Wang. Schedule your edit: A simple yet effective diffusion noise schedule for image editing. *arXiv preprint arXiv:2410.18756*, 2024. 2

[30] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3

[31] Aoyang Liu, Qingnan Fan, Shuai Qin, Hong Gu, and Yansong Tang. Lipe: Learning personalized identity prior for non-rigid image editing. *arXiv preprint arXiv:2406.17236*, 2024. 2

[32] Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024. 3

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. 3

[34] Qiang Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022. 3, 4

[35] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2022. 3, 4, 6

[36] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *CVPR*, pages 3491–3500, 2024. 3

[37] Yong Liu, Cairong Zhang, Yitong Wang, Jiahao Wang, Yujiu Yang, and Yansong Tang. Universal segmentation at arbitrary granularity with language instruction. In *CVPR*, pages 3459–3469, 2024. 3

[38] Yue Ma, Yali Wang, Yue Wu, Ziyu Lyu, Siran Chen, Xiu Li, and Yu Qiao. Visual knowledge graph for human action reasoning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4132–4141, 2022. 8

[39] Yue Ma, Xiaodong Cun, Yingqing He, Chenyang Qi, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Magicstick: Controllable video editing via control handle transformations. *arXiv preprint arXiv:2312.03047*, 2023.

[40] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024.

[41] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024.

[42] Yue Ma, Kunyu Feng, Xinhua Zhang, Hongyu Liu, David Junhao Zhang, Jinbo Xing, Yinhan Zhang, Ayden Yang, Zeyu Wang, and Qifeng Chen. Follow-your-creation: Empowering 4d creation through video inpainting. *arXiv preprint arXiv:2506.04590*, 2025.

[43] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Leqi Shen, Chenyang Qi, Jixuan Ying, Chengfei Cai, Zhifeng Li, Heung-Yeung Shum, et al. Follow-your-click: Open-domain regional image animation via motion prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6018–6026, 2025.

[44] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025. 8

[45] 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. In *ICLR*, 2022. 2

[46] 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

[47] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, pages 6038–6047, 2023. 2

[48] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 2, 4, 10, 12

[49] 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 *ICML*, pages 8748–8763, 2021. 7, 8

[50] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[52] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 2, 3, 4

[53] Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. *arXiv preprint arXiv:2410.10792*, 2024. 6, 7, 8

[54] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 7

[55] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3, 4, 6

[56] 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. 3

[57] Siao Tang, Xin Wang, Hong Chen, Chaoyu Guan, Zewen Wu, Yansong Tang, and Wenwu Zhu. Post-training quantization with progressive calibration and activation relaxing for text-to-image diffusion models. In *ECCV*, pages 404–420, 2024. 2

[58] Yoad Tewel, Rinon Gal, Dvir Samuel, Yuval Atzmon, Lior Wolf, and Gal Chechik. Add-it: Training-free object insertion in images with pretrained diffusion models. *arXiv preprint arXiv:2411.07232*, 2024. 2, 5

[59] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, pages 1921–1930, 2023. 2, 5, 12

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. page 6000–6010, 2017. 4

[61] Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. Towards accurate post-training quantization for diffusion models. In *CVPR*, pages 16026–16035, 2024. 2

[62] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *arXiv preprint arXiv:2406.08850*, 2024. 8

[63] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024. 2, 6, 7, 8

[64] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 3

[65] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *NeurIPS*, 36:15903–15935, 2023. 7, 8

[66] Sihan Xu, Yidong Huang, Jiayi Pan, Ziqiao Ma, and Joyce Chai. Inversion-free image editing with language-guided diffusion models. In *CVPR*, pages 9452–9461, 2024. 2, 5, 6

[67] Yu Xu, Fan Tang, Juan Cao, Yuxin Zhang, Xiaoyu Kong, Jintao Li, Oliver Deussen, and Tong-Yee Lee. Headrouter: A training-free image editing framework for mm-dits by adaptively routing attention heads. *arXiv preprint arXiv:2411.15034*, 2024. 2

[68] Zexuan Yan, Yue Ma, Chang Zou, Wenteng Chen, Qifeng Chen, and Linfeng Zhang. Eedit: Rethinking the spatial and temporal redundancy for efficient image editing. *arXiv preprint arXiv:2503.10270*, 2025. 8

[69] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18155–18165, 2022. 3

[70] Zhao Yang, Jiaqi Wang, Xubing Ye, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Language-aware vision transformer for referring segmentation. *TPAMI*, 2024. 3

[71] Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. Atp-llava: Adaptive token pruning for large vision language models. *arXiv preprint arXiv:2412.00447*, 2024. 3

[72] Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, Ying Shan, and Yansong Tang. Voco-llama: Towards vision compression with large language models. *arXiv preprint arXiv:2406.12275*, 2024.

[73] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024. 3

[74] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 7

[75] Yinhan Zhang, Yue Ma, Bingyuan Wang, Qifeng Chen, and Zeyu Wang. Magiccolor: Multi-instance sketch colorization. *arXiv preprint arXiv:2503.16948*, 2025. 8

[76] Chenyang Zhu, Kai Li, Yue Ma, Longxiang Tang, Chengyu Fang, Chubin Chen, Qifeng Chen, and Xiu Li. Instantswap: Fast customized concept swapping across sharp shape differences. *arXiv preprint arXiv:2412.01197*, 2024. 8

[77] Yixuan Zhu, Wenliang Zhao, Ao Li, Yansong Tang, Jie Zhou, and Jiwen Lu. Flowie: Efficient image enhancement via rectified flow. In *CVPR*, pages 13–22, 2024. 2

[78] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *ECCV*, pages 195–211, 2024. 2, 3