InverFill: One-Step Inversion for Enhanced Few-Step Diffusion Inpainting

Supplementary Material

We first present ablations on the loss weights in Sec. 7. Sec. 8 compares our method with other regularization techniques, while Sec. 10 evaluates alternative inversion methods. Additional ablations for our proposed components are in Sec. 9. Sec. 15 includes qualitative comparisons.

Note: All experiments in both the main paper and supplementary use 1024^2 resolution. For all supplementary results, we use BrushBench [5] with its original captions.

7. Loss Weight Ablations

We evaluate the impact of reconstruction weights λ_{noise} and λ_{image} in $\mathcal{L}_{\text{recons}}$ (Sec. 4.3), along with the Gaussian regularization λ_{reg} (Sec. 4.5) and adversarial weights λ_{adv} (Sec. 4.6), using SANA-Sprint 0.6B [2]. All experiments use the Re-Blending operation (Sec. 4.4) during training and inference. For LADD adversarial loss, the discriminator learning rate is set to 1×10^{-6}. Detailed results are provided in Tab. 4.

Table 4. Ablation study on key hyperparameters for each component. The best setting from each block is propagated to the next.

Method	λ_{noise}	λ_{image}	λ_{reg}	λ_{adv}	IR$_{\times 10}$↑	HPS$_{\times 10^2}$↑	AS↑	CLIP↑
Sec. 4.3	2.0	1.0	0	0	11.09	**26.69**	6.04	27.10
	1.0	2.0	0	0	10.91	26.67	6.06	27.05
	1.0	**1.0**	0	0	**11.11**	26.68	**6.08**	**27.13**
Sec. 4.5	1.0	1.0	0.25	0	11.12	26.55	6.09	27.13
	1.0	1.0	**0.5**	0	**11.40**	**27.22**	**6.12**	**27.15**
	1.0	1.0	1.0	0	11.36	26.58	6.10	27.14
	1.0	1.0	2.0	0	11.03	26.56	6.09	27.17
Sec. 4.6	1.0	1.0	0.5	0.25	11.57	27.36	6.14	27.16
	1.0	1.0	0.5	**0.5**	**11.65**	**27.93**	**6.15**	**27.17**
	1.0	1.0	0.5	1.0	11.60	27.61	6.15	27.17

Tab. 4 summarizes the ablation results on the loss-weight components. Based on this study, we use the final weights $\lambda_{\text{noise}} = 1.0$, $\lambda_{\text{image}} = 1.0$, $\lambda_{\text{reg}} = 0.5$, and $\lambda_{\text{adv}} = 0.5$ for all experiments reported in Tabs. 1 and 3.

8. Comparison with Regularization Loss in SwiftEdit

We perform an ablation to compare our regularization loss $\mathcal{L}_{reg}$ with the Score Distillation Sampling loss $\mathcal{L}_{\text{SDS}}$ used in SwiftEdit [8]. As shown in Tab. 5, applying $\mathcal{L}_{reg}$ consistently outperforms $\mathcal{L}_{\text{SDS}}$ across all metrics (IR, HPS, AS, and CLIP), demonstrating its effectiveness in preserving image fidelity. Fig. 8 illustrates the qualitative difference between the two losses. With the SDS-based loss, the reconstruction collapses, as the inverted noise is over-regularized and loses the semantic structure of the original image, producing blurry and unrecognizable results. In contrast, our Gaussian regularization loss $\mathcal{L}_{reg}$ preserves the semantic content and enables high-fidelity reconstruction from the inverted noise.

Table 5. Quantitative comparison between the SDS loss $\mathcal{L}_{\text{SDS}}$ and our Gaussian regularization loss $\mathcal{L}_{reg}$. For a fair evaluation, both methods are tested on SANA-Sprint 0.6B using 2 NFEs.

Method	IR$_{\times 10}$↑	HPS$_{\times 10^2}$↑	AS↑	CLIP↑
$\mathcal{L}_{\text{SDS}}$ [8]	11.18	26.50	6.10	27.12
$\mathcal{L}_{reg}$ (Ours)	**11.40**	**27.22**	**6.12**	**27.15**

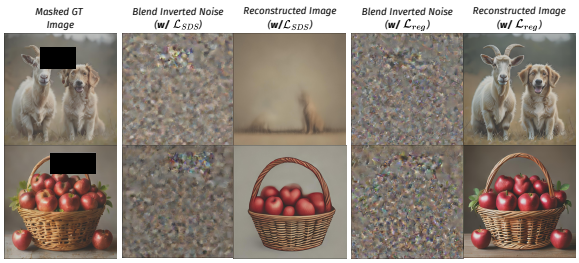

Figure 8. **Qualitative comparison between our proposed regularization loss ($\mathcal{L}_{reg}$) and the Score Distillation Sampling (SDS) loss ($\mathcal{L}_{\text{SDS}}$) from SwiftEdit [8].** This visualization shows that our $\mathcal{L}_{reg}$ is crucial for preserving the original image content, while using $\mathcal{L}_{\text{SDS}}$ leads to significant information loss and poor reconstruction.

9. Ablation of Proposed Components

To better understand the contribution of each part in our framework, we conducted an ablation study on both SANA-Sprint 0.6B (Tab. 7) and SDXL-Turbo (Tab. 8). We established a baseline for comparison by training a model with the reconstruction loss from Sec. 4.3, using the masked image as input. From this starting point, we then incrementally added our proposed components: Re-Blending (Sec. 4.4), Gaussian Regularization (Sec. 4.5), and the LADD adversarial loss (Sec. 4.6).

Our results show that each component contributes incremental gains in performance. As shown in Tab. 7, introducing the Re-Blending operation increases the IR score from 7.93 to 11.11. The further addition of Gaussian Regularization expands this improvement, and incorporating the LADD adversarial loss leads to the highest scores, with an IR of 11.65 and an HPS of 27.93. A similar pattern of improvement is also noted in the experiments with SDXL-Turbo (Tab. 8). This evaluation suggests that all three components contribute effectively, collectively leading to the

Method	IR$_{\times 10}$ ↑	HPS$_{\times 10^2}$	AS ↑	CLIP↑	**Runtime** (seconds)↓
DDIMInv (**w/o** Blending) (50 steps) + SDXL-Turbo	4.10	22.95	5.30	26.40	4.18
DDIMInv (**w/** Blending) (50 steps) + SDXL-Turbo	12.11	28.21	6.04	27.32	4.32
InverFill (Ours) + SDXL-Turbo	**12.38**	**28.44**	**6.08**	**27.67**	**0.74**

Table 6. Quantitative comparison of one-step InverFill versus the 50-step DDIM inversion baseline on BrushBench. For a fair comparison, we run SDXL-Turbo with 4 NFEs.

performance of the full InverFill model.

Table 7. Ablation of components on SANA-Sprint 0.6B (2 NFEs).

Method	IR$_{\times 10}$↑	HPS$_{\times 10^2}$↑	AS↑	CLIP↑
Baseline [8]	7.93	24.79	5.96	26.40
InverFill				
+ Re-Blending (Sec. 4.4)	11.11	26.68	6.08	27.13
+ Gaussian Reg. (Sec. 4.5)	11.40	27.22	6.12	27.15
+ LADD (Sec. 4.6)	**11.65**	**27.93**	**6.15**	**27.17**

Table 8. Ablation of components on SDXL-Turbo (4 NFEs).

Method	IR$_{\times 10}$↑	HPS$_{\times 10^2}$↑	AS↑	CLIP↑
Baseline [8]	10.64	26.46	6.03	26.56
InverFill				
+ Re-Blending (Sec. 4.4)	11.33	27.18	6.03	27.16
+ Gaussian Reg. (Sec. 4.5)	12.14	28.14	6.06	27.57
+ LADD (Sec. 4.6)	**12.38**	**28.44**	**6.08**	**27.67**

10. Other Inversion Approaches

We quantitatively compare InverFill with a 50-step DDIM inversion process, using SDXL for inversion and SDXL-Turbo blended sampling for inpainting. Based on Fig. 9, directly applying DDIM inversion to a masked image fails to encode the masked regions, producing smooth, gray, null-like structures in those areas. This loss of content significantly degrades performance, as reflected in the low scores reported in the first row of Tab. 6.

Next, we apply our proposed Re-Blending operation (Sec. 4.4) to the DDIM-inverted noise. While this fills the previously null-like regions, the resulting model (Row 2 in Tab. 6) still struggles with scene harmonization. In contrast, our one-step InverFill method (Row 3) achieves higher scores across all metrics and is significantly more efficient, running in just **0.74 seconds**, nearly **six times faster** than the 50-step DDIM process (4.32 seconds). Combined with the improved qualitative harmonization in Fig. 10, these results demonstrate that InverFill is both substantially more effective and practical.

11. Additional Experiments

We evaluate InverFill on FFHQ [6] and DIV2K [1] to assess robustness across diverse mask configurations and standard

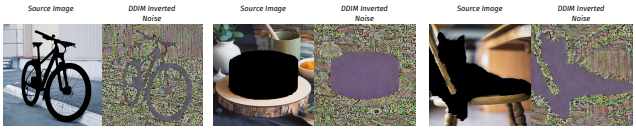

Figure 9. Visualization of DDIM inversion results. The masked regions are not encoded, producing smooth, null-like areas in the inverted noise and causing loss of content.

Figure 10. Qualitative comparison between InverFill and DDIM inversion. InverFill achieves substantially better scene harmonization and semantic consistency.

benchmarks, with results reported in Tab. 9. For all evaluations, we use the same checkpoints as in Sec. 5 without any modification or fine-tuning.

Datasets. For FFHQ, we sample 10K images. For DIV2K, we use 900 images from the training and validation sets. Following the same settings in Sec. 5.5, prompts are generated using Qwen-3 [12].

Mask Settings. For both FFHQ and DIV2K, we adopt LaMa's [11] strategy with polygonal thick- and thin-stroke masks, and additionally include rectangular masks covering half of the image. Masks are randomly sampled from these configurations to ensure a diverse evaluation.

Additional Metrics. In addition to perceptual quality met-

Table 9. **Quantitative results on FFHQ, DIV2K, and BrushBench.** **Red**, **Blue**, and **Black** denote scores on **FFHQ**, **DIV2K**, and **BrushBench**, respectively.

Method	NFEs	FFHQ / DIV2K / BrushBench						
		FID↓	IR$_{\times 10}$↑	HPS$_{\times 10^2}$↑	AS↑	CLIP↑	LPIPS↓	SSIM↑
SANA-Sprint 0.6B	2	27.12	2.81 / 5.22	22.42 / 26.78	5.12 / 5.77	23.55 / 28.26	0.184 / 0.193 / 0.144	0.704 / 0.572 / 0.769
SANA-Sprint 0.6B + InverFill	2	**26.53**	**5.27** / **5.87**	**23.26** / **27.17**	**5.31** / **5.89**	**23.65** / **28.43**	**0.172** / **0.182** / **0.138**	**0.719** / **0.575** / **0.771**
SANA-Sprint 0.6B	4	27.32	2.66 / 5.25	22.50 / 26.83	5.17 / 5.79	23.84 / 28.31	0.184 / 0.192 / 0.140	0.706 / 0.573 / 0.774
SANA-Sprint 0.6B + InverFill	4	**26.42**	**5.27** / **5.83**	**23.32** / **27.15**	**5.37** / **5.92**	**23.88** / **28.38**	**0.169** / **0.181** / **0.134**	**0.708** / **0.574** / 0.774
SDXL Turbo	4	26.32	7.37 / 4.71	25.73 / 26.81	5.67 / 5.92	25.24 / 28.21	0.269 / 0.292 / 0.139	0.626 / 0.454 / 0.813
SDXL Turbo + InverFill	4	**25.90**	**8.35** / **5.27**	**26.14** / **27.03**	**5.76** / **5.95**	**25.29** / **28.25**	**0.262** / **0.287** / **0.133**	**0.655** / **0.455** / **0.815**
SDXL Turbo + BrushNet	4	25.55	7.86 / 5.11	25.05 / 26.05	5.53 / 5.76	24.72 / **28.41**	**0.204** / 0.469 / 0.185	**0.728** / 0.292 / 0.755
SDXL Turbo + BrushNet + InverFill	4	**25.49**	**7.91** / **5.17**	**25.17** / **26.18**	**5.55** / **5.79**	**24.85** / 28.39	0.206 / 0.469 / **0.178**	0.727 / **0.293** / **0.757**

rics, we report LPIPS and SSIM to assess consistency, including results from the BrushBench evaluation. For FFHQ, we additionally report FID [4].

12. Analysis of the Inversion Effect

We analyze the effect of the inversion network to explain why initializing from well-aligned noise yields more coherent and consistent outputs. Our intuition is that such noise encodes the blending trajectory and preserves background information, thereby enabling smoother blending during the denoising process.

To further validate this observation, we compute LPIPS between the predicted x_0 in background regions at intermediate timesteps and the input image, and report the results in Fig. 11. We observe that initialization with well-aligned noise consistently yields significantly lower LPIPS than random initialization, supporting our hypothesis. Moreover, Fig. 11 provides insight into the effectiveness of the Gaussian regularization loss: the Jensen–Shannon divergence (JSD) with respect to the Gaussian distribution is substantially reduced when this regularization is applied, leading to better-aligned latent noise while also satisfying the required Gaussian distribution for diffusion models, and consequently yielding stable and coherent reconstructions.

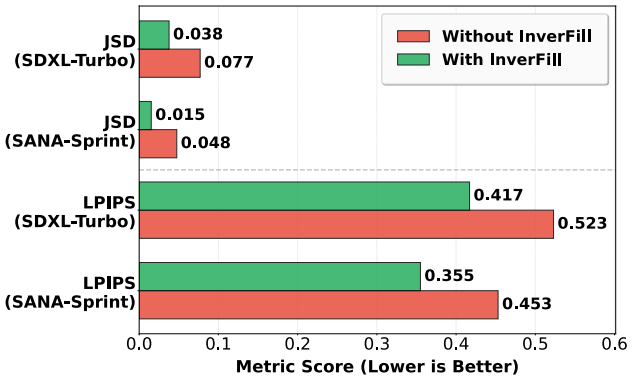

Figure 11. **Quantitative analysis of inversion effects.** We report LPIPS in background regions at intermediate timesteps and JSD with respect to the Gaussian prior. Lower values indicate better alignment. Red bars denote results without InverFill, while Green bars denote results with InverFill.

13. Failure Cases

We report representative failure cases in Fig. 12. Overall, the main limitation of our method stems from color inconsistencies between the inpainted region and the background.

Figure 12. Representative failure cases of our method. While InverFill improves overall coherence, it may produce color inconsistencies between the inpainted regions and the background.

14. Societal Impacts

Our work aims to provide a practical tool for creative professionals, facilitating tasks such as photo restoration and object removal. We acknowledge that realistic image manipulation technologies can be misused to generate deceptive content. To mitigate such risks, we advocate for the parallel development of detection methods [3, 7, 9] for AI-manipulated media and encourage the responsible use of these technologies.

15. More Qualitative Results

To provide a comprehensive visual comparison of InverFill, Figs. 13 and 14 present an expanded set of qualitative results, further illustrating the improvements in coherence and background harmonization highlighted in our work.

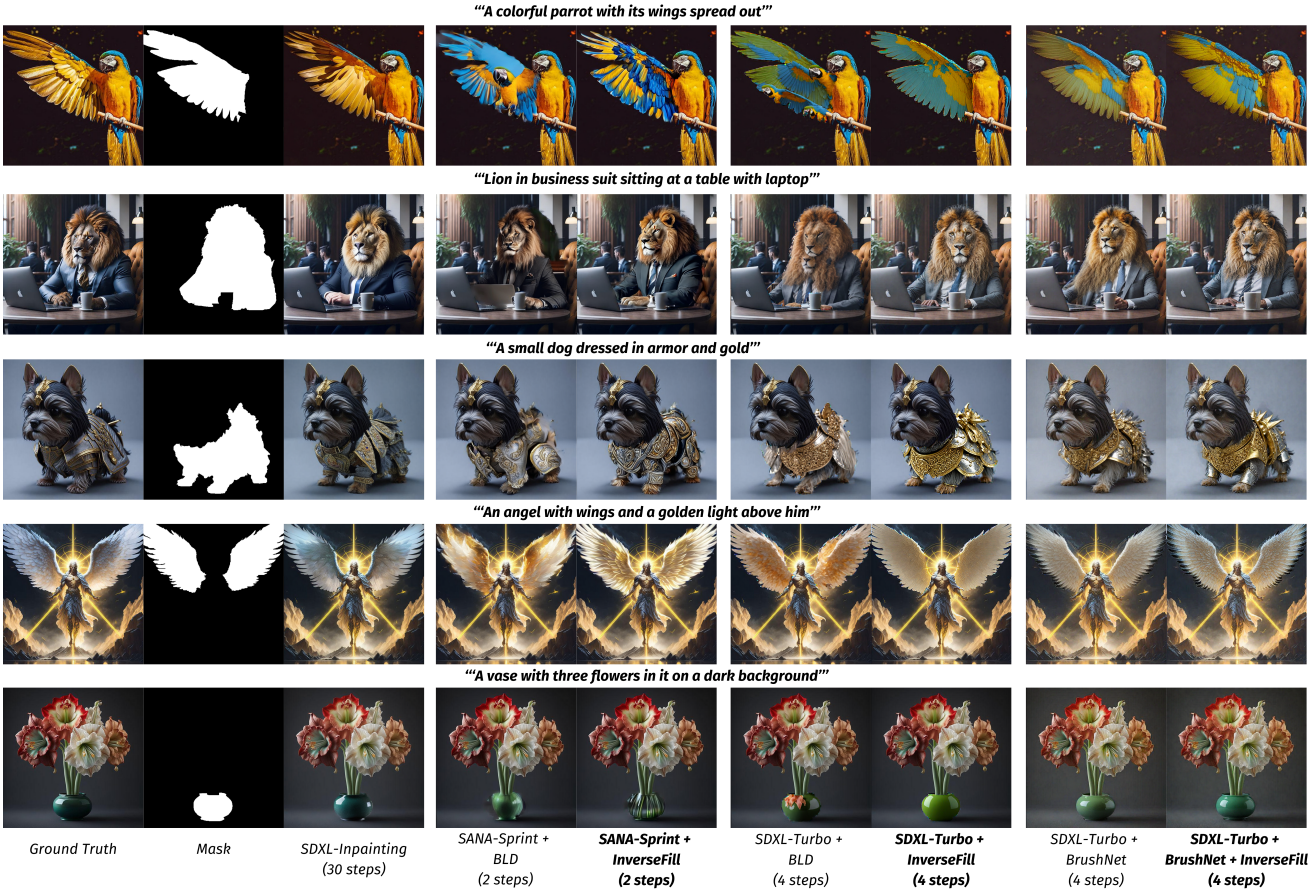

Figure 13. More qualitative comparison on BrushBench *(Zoom in for best view)*

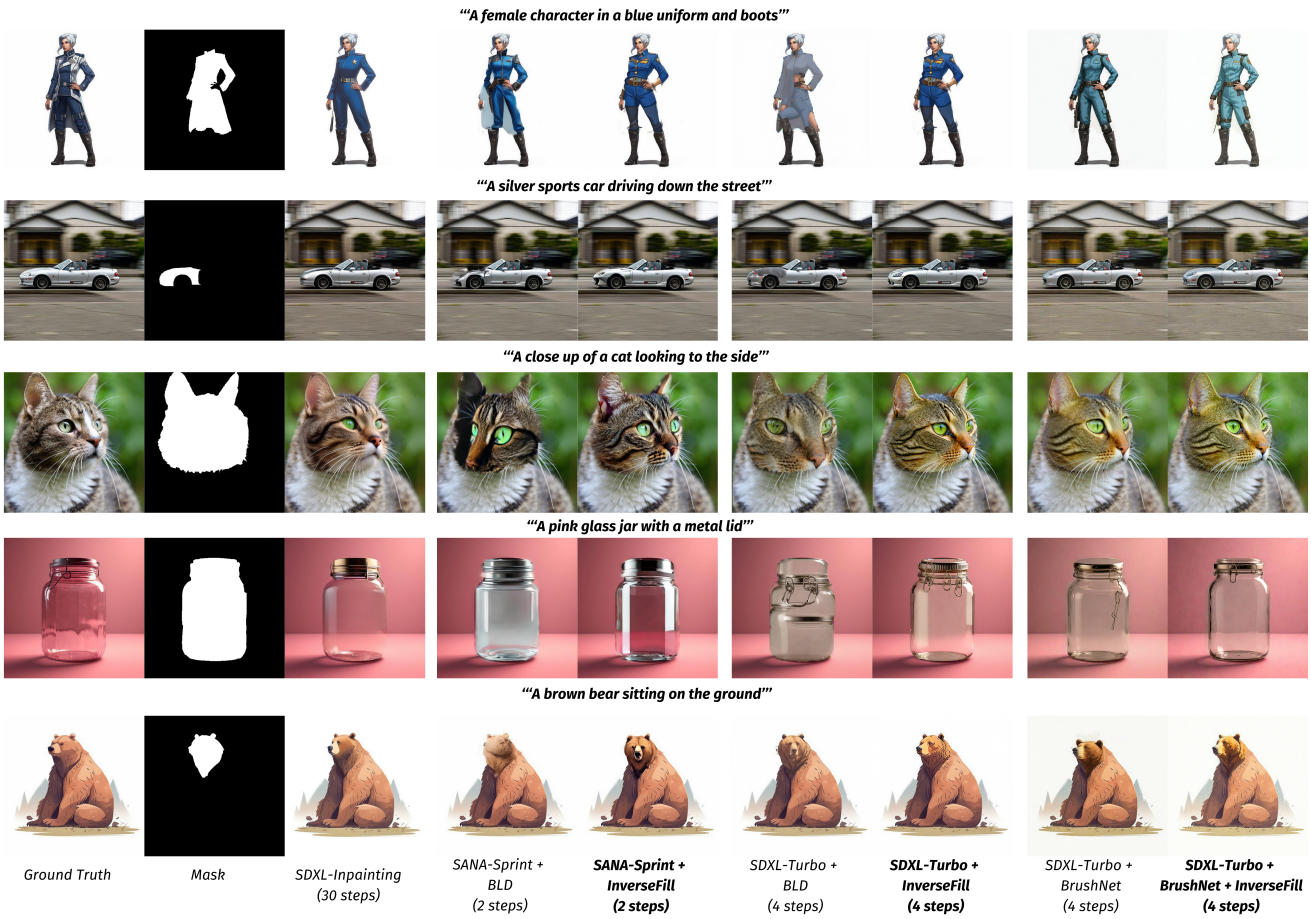

Figure 14. More qualitative comparison on BrushBench *(Zoom in for best view)*

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, 2017. 2

[2] Junsong Chen, Shuchen Xue, Yuyang Zhao, Jincheng Yu, Sayak Paul, Junyu Chen, Han Cai, Enze Xie, and Song Han. Sana-sprint: One-step diffusion with continuous-time consistency distillation. *CoRR*, 2025. 2, 3, 6, 8, 1

[3] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *ICCV*, pages 2426–2436, 2023. 3

[4] 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. *NeurIPS*, 2017. 3

[5] 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 *European Conference on Computer Vision*, pages 150–168, 2024. 1, 3, 7, 8

[6] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2

[7] Kien Nguyen, Anh Tran, and Cuong Pham. Suma: A subspace mapping approach for robust and effective concept erasure in text-to-image diffusion models. In *ICCV*, pages 19587–19596, 2025. 3

[8] Trong-Tung Nguyen, Quang Nguyen, Khoi Nguyen, Anh Tran, and Cuong Pham. Swiftedit: Lightning fast text-guided image editing via one-step diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21492–21501, 2025. 2, 3, 5, 1

[9] Viet Nguyen and Vishal M Patel. Cgce: Classifier-guided concept erasure in generative models. *arXiv preprint arXiv:2511.05865*, 2025. 3

[10] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 1, 2, 3, 4, 8

[11] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, pages 2149–2159, 2022. 2

[12] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 8, 2