Supplementary Material for
Semantic Context Matters: Improving Conditioning for Autoregressive Models

1. More Discussion

Decoding-stage Injection vs. Prefilling-stage condition. Conditioning strategies for autoregressive (AR) models fall into two categories: decoding-stage injection and prefilling-stage conditioning. We argue that the latter provides a more suitable conditioning mechanism for instruction editing and controllable generation, while remaining naturally compatible with both next-token and next-set AR paradigms.

For controllable generation, decoding-stage methods [5, 8, 19, 21] like ControlAR [5] and CAR [21] inject control signals into intermediate layers, enabling fine-grained spatial alignment. However, this tight coupling often leads to training instability and overfitting to local structures, resulting in blurry textures and reduced realism. In contrast, prefilling-based methods such as EditAR [10] and our SCAR prepend control features as conditioning tokens before decoding. Though offering slightly weaker pixel-level control, this design leads to sharper, more natural outputs and better generalization across control types. More importantly, for instruction editing, decoding-stage injection fails to align generation with high-level semantics, often producing overconstrained or inconsistent results. It also conflicts with unified multimodal models (UMM) [1, 2, 6, 7, 9, 12, 15, 18], as spatial injection disrupts the shared AR generation. Prefilling-stage conditioning avoids this issue, integrates cleanly with UMM, and better preserves semantic intent during generation.

Semantic Prefix vs. VQ Prefix. Table 7 reports the quantitative comparison between semantically aligned DINO tokens and standard VQ tokens under two compression settings (k^2=1× and 4×) on MultiGen-20M [14]. Across both settings, DINO tokens consistently outperform VQ tokens, underscoring the effectiveness of our semantic prefix for autoregressive editing. Without the compression (1× compression) setting, DINO tokens improve SSIM by +1.96 and reduce FID by 0.91, indicating better perceptual quality and semantic coherence. Under the 4× compression setting, the improvement becomes even more pronounced: SSIM increases by +3.21, and FID drops by 2.83. These results demonstrate that our method not only improves generation quality in standard setups but also maintains strong performance under token compression.

Table 7. Comparison between the semantical DINO token and the VQ token under different compression ratios. All models are trained for 1 epoch. Values in parentheses indicate relative improvement over VQ tokens.

Prefix Condition	Compress Ratio k^2	HED	
		FID↓	SSIM↑
VQ token	1×	10.20	79.99
DINO token	1×	9.29 (-0.91)	81.95 (+1.96)
VQ token	4×	12.26	78.55
DINO token	4×	9.43 (-2.83)	81.76 (+3.21)

Discussion on Semantic Alignment. Recent work increasingly suggests that enforcing semantic representation alignment can lead to more effective learning in generative models. In diffusion models, REPA [22] aligns internal features with pretrained semantic encoders to stabilize training and improve generation. REG [17] takes a different path by injecting discriminative semantics directly via spatial concatenation with a class token, improving quality and convergence with minimal overhead. VA-VAE [20] enhances latent interpretability via representation alignment in VAEs. MaskDiT [23] enforces semantic consistency through masked reconstruction or auxiliary decoders. Multi-stage methods [4, 13] leverage pretrained representations as intermediate maps, while USP [3] aligns masked latents in a shared VAE space to unify generation and understanding.

While these strategies have shown strong results in diffusion-based models, extending semantic alignment to autoregressive (AR) frameworks for editing presents unique challenges. EditAR [10] attempts this by distilling supervision from DINO features onto the hidden states of generated VQ tokens. However, as this supervision is applied post-generation, it lacks explicit semantic flow during decoding.

In contrast, SCAR aligns semantic features at the prefilling stage, by matching conditional representations from the source image with a pretrained visual space (e.g., DINO [11]) before decoding begins. This design fits naturally with the AR model's causal structure, introducing dense supervision without interfering with token-level predictions. As a result, the model internalizes the semantic correspondence early and propagates it consistently

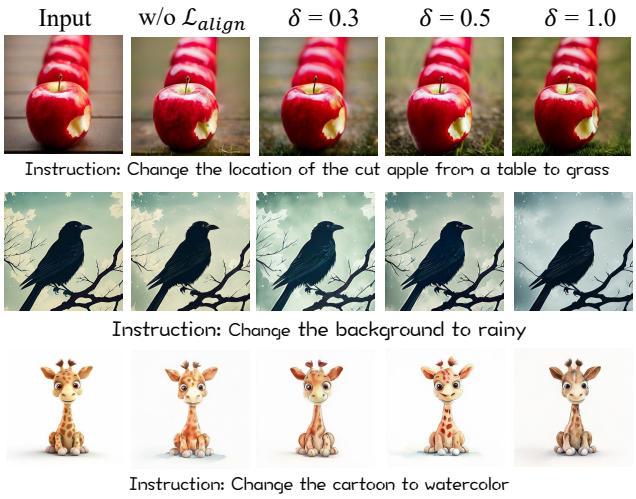

Figure 8. Additional visualizations on Semantic Alignment Guidance $\mathcal{L}_{align}$ in Equation (8).

throughout generation. Compared to global alignment strategies in diffusion or post-hoc distillation in EditAR, SCAR provides a more direct and effective mechanism for integrating conditional guidance, especially in instruction-based editing where semantic consistency across text, condition, and output is critical.

2. More Visualization

2.1. About Semantic Alignment Guidance

As shown in Figure 8, without Semantic Alignment Guidance, the model may sometimes produce incomplete or ambiguous edits. Semantic Alignment Guidance mitigates this by strengthening the correspondence between input conditions and target semantics. In the apple relocation example (top row), both the *w/o* $\mathcal{L}_{\text{align}}$ and δ=0.3 results retain table textures, indicating insufficient background change. In contrast, δ=0.5 successfully introduces grass while preserving the foreground. However, δ=1.0 causes a global green tint, suggesting semantic leakage. In the background editing task (middle row), δ=1.0 enhances the rainy effect but removes fine details such as leaves, while δ=0.5 better balances visual accuracy and structure. For stylization (bottom row), larger δ values improve the watercolor effect, but δ=1.0 leads to facial distortion and oversaturation.

Overall, moderate alignment strength improves consistency and visual quality, whereas excessive supervision may cause artifacts or semantic drift.

2.2. Instruction Editing

As shown in Figure 9, we present additional visualizations for instruction editing, further demonstrating the effectiveness of our proposed SCAR.

2.3. Controllable Generation

C2I Controllable Generation by SCAR-Uni. In C2I Controllable Generation, we focus on the results of SCAR-Uni, a unified method that supports diverse control conditions using the same model. As shown in Figure 10, SCAR-Uni generates high-quality images while maintaining strong controllability across diverse input types.

T2I Controllable Generation. Figure 11 presents additional visualizations for T2I controllable generation. The results further demonstrate the strong performance and generalization ability of our proposed SCAR.

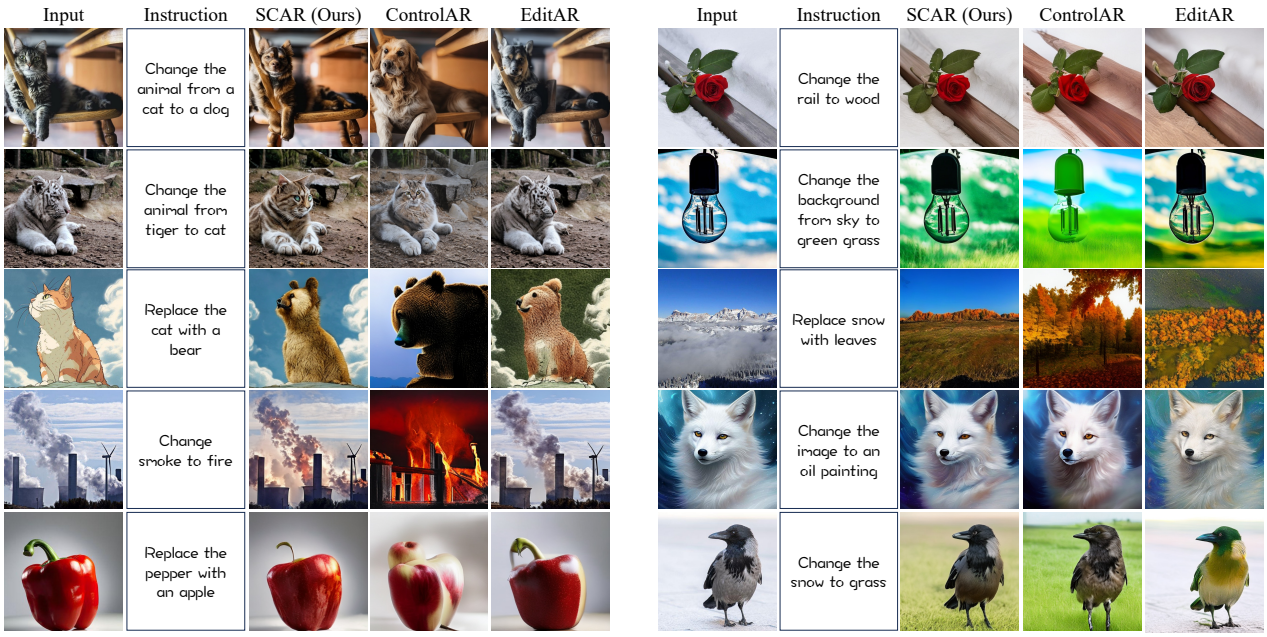

Figure 9. Additional visualizations of instruction editing results. SCAR (Ours) produces more faithful and semantically consistent edits than ControlAR [5] and EditAR [10], with all methods using the same LlamaGen-XL [16]. All visualizations are generated at a resolution of 512×512.

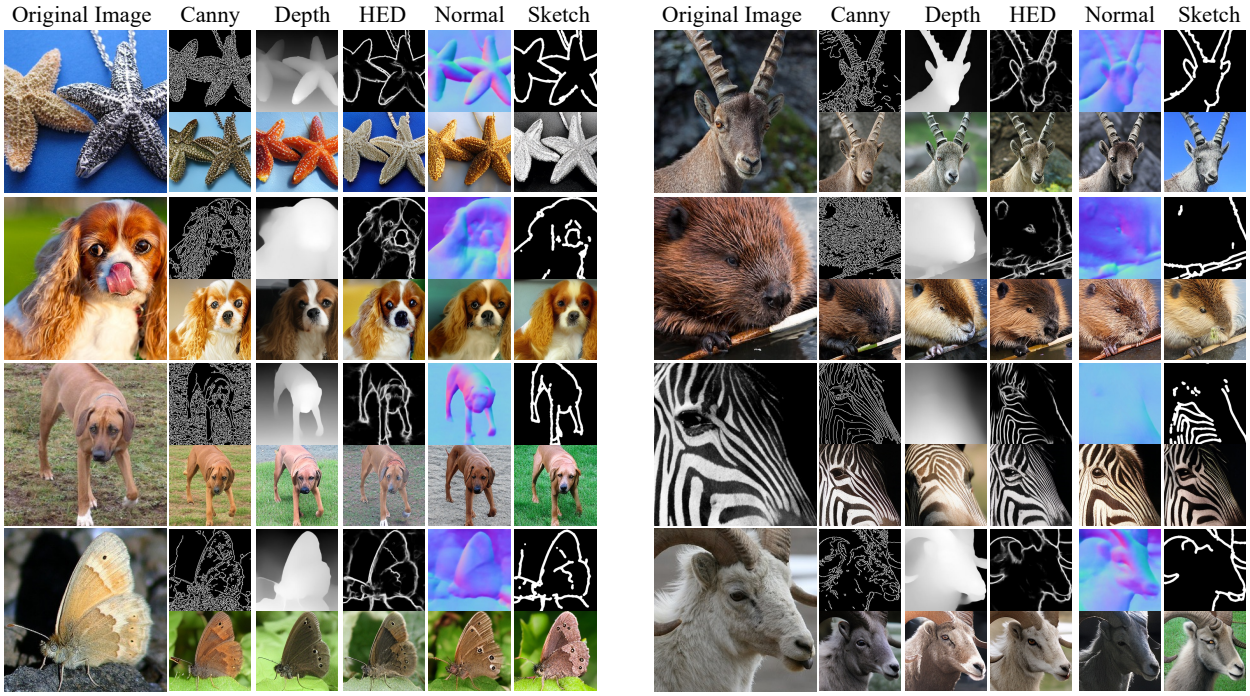

Figure 10. Additional visualizations of C2I Controllable Generation by SCAR-Uni based on LlamaGen-L. We adopt five different control conditions: Canny, Depth, HED, Normal, and Sketch. All visualizations are generated at a resolution of 256×256.

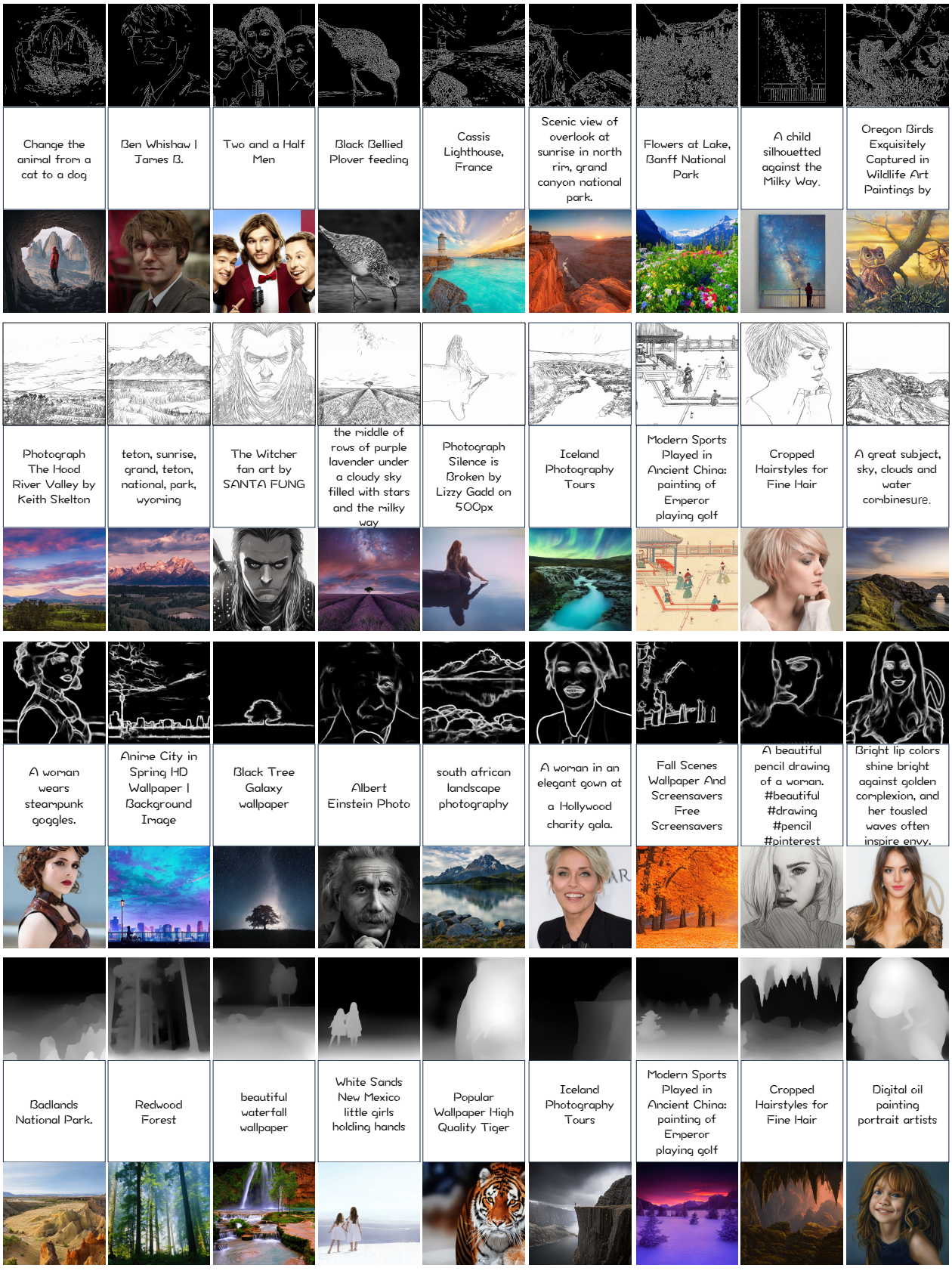

Figure 11. Additional visualizations of T2I Controllable Generation by SCAR based on LlamaGen-XL. To demonstrate the controllability and generalization ability of our SCAR, we present results under four control conditions: Canny, Depth, HED, and Lineart. All images are generated at a resolution of 512×512.

References

[1] Ruichuan An, Sihan Yang, Ming Lu, Renrui Zhang, Kai Zeng, Yulin Luo, Jiajun Cao, Hao Liang, Ying Chen, Qi She, et al. Mc-llava: Multi-concept personalized vision-language model. *arXiv preprint arXiv:2411.11706*, 2024. 1

[2] Ruichuan An, Sihan Yang, Renrui Zhang, Zijun Shen, Ming Lu, Gaole Dai, Hao Liang, Ziyu Guo, Shilin Yan, Yulin Luo, et al. Unictokens: Boosting personalized understanding and generation via unified concept tokens. *arXiv preprint arXiv:2505.14671*, 2025. 1

[3] Xiangxiang Chu, Renda Li, and Yong Wang. Usp: Unified self-supervised pretraining for image generation and understanding. *arXiv preprint arXiv:2503.06132*, 2025. 1

[4] Tianhong Li, Dina Katabi, and Kaiming He. Return of unconditional generation: A self-supervised representation generation method. *Advances in Neural Information Processing Systems*, 37:125441–125468, 2024. 1

[5] Zongming Li, Tianheng Cheng, Shoufa Chen, Peize Sun, Haocheng Shen, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Controlar: Controllable image generation with autoregressive models. *arXiv preprint arXiv:2410.02705*, 2024. 1, 3

[6] Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *arXiv preprint arXiv:2403.20271*, 2024. 1

[7] Weifeng Lin, Xinyu Wei, Ruichuan An, Tianhe Ren, Tingwei Chen, Renrui Zhang, Ziyu Guo, Wentao Zhang, Lei Zhang, and Hongsheng Li. Perceive anything: Recognize, explain, caption, and segment anything in images and videos, 2025. 1

[8] Keli Liu, Zhendong Wang, Wengang Zhou, Shaodong Xu, Ruixiao Dong, and Houqiang Li. Scaleweaver: Weaving efficient controllable t2i generation with multi-scale reference attention. *arXiv preprint arXiv:2510.14882*, 2025. 1

[9] Yulin Luo, Ruichuan An, Bocheng Zou, Yiming Tang, Jiaming Liu, and Shanghang Zhang. Llm as dataset analyst: Subpopulation structure discovery with large language model. In *European Conference on Computer Vision*, pages 235–252. Springer, 2024. 1

[10] Jiteng Mu, Nuno Vasconcelos, and Xiaolong Wang. Editar: Unified conditional generation with autoregressive models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7899–7909, 2025. 1, 3

[11] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1

[12] Long Peng, Yang Cao, Renjing Pei, Wenbo Li, Jiaming Guo, Xueyang Fu, Yang Wang, and Zheng-Jun Zha. Efficient real-world image super-resolution via adaptive directional gradient convolution. *arXiv preprint arXiv:2405.07023*, 2024. 1

[13] Pablo Pernias, Dominic Rampas, Mats L Richter, Christopher J Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. *arXiv preprint arXiv:2306.00637*, 2023. 1

[14] Can Qin, Shu Zhang, Ning Yu, Yihao Feng, Xinyi Yang, Yingbo Zhou, Huan Wang, Juan Carlos Niebles, Caiming Xiong, Silvio Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 1

[15] Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. Ultrapixel: Advancing ultra high-resolution image synthesis to new peaks. *Advances in Neural Information Processing Systems*, 37:111131–111171, 2024. 1

[16] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 3

[17] Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, Jian Yang, et al. Representation entanglement for generation: Training diffusion transformers is much easier than you think. *arXiv preprint arXiv:2507.01467*, 2025. 1

[18] Hao Xu, Long Peng, Shezheng Song, Xiaodong Liu, Ma Jun, Shasha Li, Jie Yu, and Xiaoguang Mao. Camel: Energy-aware llm inference on resource-constrained devices. *arXiv preprint arXiv:2508.09173*, 2025. 1

[19] Ryan Xu, Dongyang Jin, Yancheng Bai, Rui Lan, Xu Duan, Lei Sun, and Xiangxiang Chu. Scalar: Scale-wise controllable visual autoregressive learning. *arXiv preprint arXiv:2507.19946*, 2025. 1

[20] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15703–15712, 2025. 1

[21] Ziyu Yao, Jialin Li, Yifeng Zhou, Yong Liu, Xi Jiang, Chengjie Wang, Feng Zheng, Yuexian Zou, and Lei Li. Car: Controllable autoregressive modeling for visual generation. *arXiv preprint arXiv:2410.04671*, 2024. 1

[22] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *The Thirteenth International Conference on Learning Representations*. 1

[23] Hongkai Zheng, Weili Nie, Arash Vahdat, and Anima Anandkumar. Fast training of diffusion models with masked transformers. *arXiv preprint arXiv:2306.09305*, 2023. 1