Beyond Next-Token: Next-X Prediction for Autoregressive Visual Generation

Supplementary Material

Appendix

The supplementary material includes the following additional information:

- Sec. A details the hyper-parameters used for xAR.
- Sec. B provides a comprehensive speed comparison.
- Sec. C discusses the limitations and future directions.
- Sec. D presents visualization samples generated by xAR.

A. Hyper-parameters for xAR

We list the detailed training and inference hyper-parameters in Tab. 1.

config	value
optimizer	AdamW [1, 3]
optimizer momentum	(0.9, 0.96)
weight decay	0.02
batch size	2048
learning rate schedule	cosine decay
peak learning rate	4e-4
ending learning rate	1e-5
total epochs	800
warmup epochs	100
dropout rate	0.1
attn dropout rate	0.1
class label dropout rate	0.1
inference mode	SDE
inference steps	50

Table 1. **Detailed Hyper-parameters of xAR Models.**

B. Speed Comparison.

We compare xAR with diffusion-, flow matching-, and autoregressive-based models in Tab. 2. Our most lightweight variant, xAR-B (172M), outperforms DiT-XL (diffusion-based), SiT-XL (flow matching-based), and MAR (autoregressive-based), while achieving a 20× speedup (9.8 *vs.* 0.5 images/sec). Additionally, xAR-L surpasses the recent state-of-the-art model REPA, running 5.3× faster (3.2 *vs.* 0.6 images/sec). Finally, our largest model, xAR-H, achieves 1.24 FID on ImageNet-256, setting a new state-of-the-art, while still running 2.2× faster than REPA.

method	type	#params	FID↓	steps	images/sec
DiT-XL/2 [5]	Diff.	675M	2.27	250	0.5
SiT-XL/2 [4]	Flow.	675M	2.02	250	0.5
MAR-L [2]	AR	479M	1.78	256	0.5
xAR-B	xAR	172M	1.72	50	9.8
MAR-H [2]	MAR	943M	1.55	256	0.3
REPA [6]	Flow.	675M	1.42	250	0.6
xAR-L	xAR	608M	1.28	50	3.2
xAR-H	xAR	1.1B	1.24	50	1.3

Table 2. **Sampling Throughput Comparison.** Throughputs are evaluated as samples generated per second on a single A100 based on their official codebases.

C. Discussion and Limitations

Our empirical evaluations indicate that a square 8×8 cell configuration achieves the best performance, with no noticeable difference when using rectangular cells (*e.g.*, $k/2 \times 2k$ or $2k \times k/2$), which introduce additional complexity without clear benefits. Given that different regions in an image contain varying levels of semantic information (*e.g.*, dense object areas *vs.* uniform sky regions), future research could explore whether dynamically shaped prediction entities provide additional benefits. However, in this work, we adopt a simple yet effective square cell design, demonstrating state-of-the-art results on the challenging ImageNet generation benchmark.

D. Visualization of Generated Samples

Additional visualization results generated by xAR-H are provided from Fig. 1 to Fig. 9.

Figure 1. **Generated Samples from xAR.** xAR is able to generate high-fidelity American eagle (22) images.

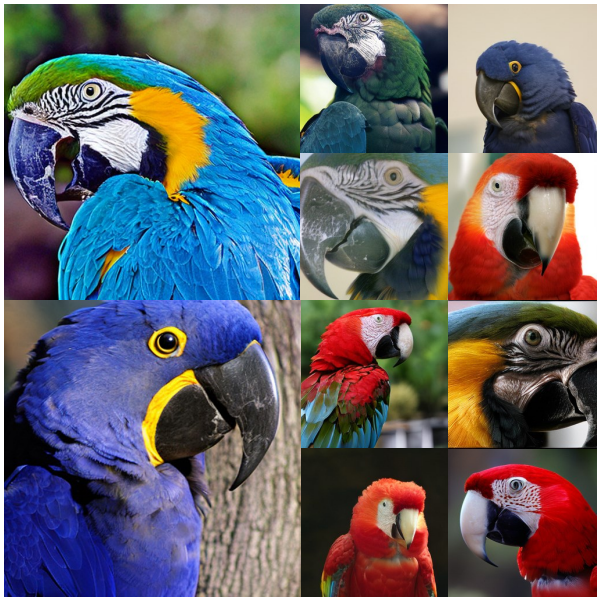

Figure 2. **Generated Samples from xAR.** xAR is able to generate high-fidelity macaw (88) images.

Figure 4. **Generated Samples from xAR.** xAR is able to generate high-fidelity otter (360) images.

Figure 3. **Generated Samples from xAR.** xAR is able to generate high-fidelity golden retriever (207) images.

Figure 5. **Generated Samples from xAR.** xAR is able to generate high-fidelity lesser panda (387) images.

Figure 6. **Generated Samples from xAR.** xAR is able to generate high-fidelity coral reef (973) images.

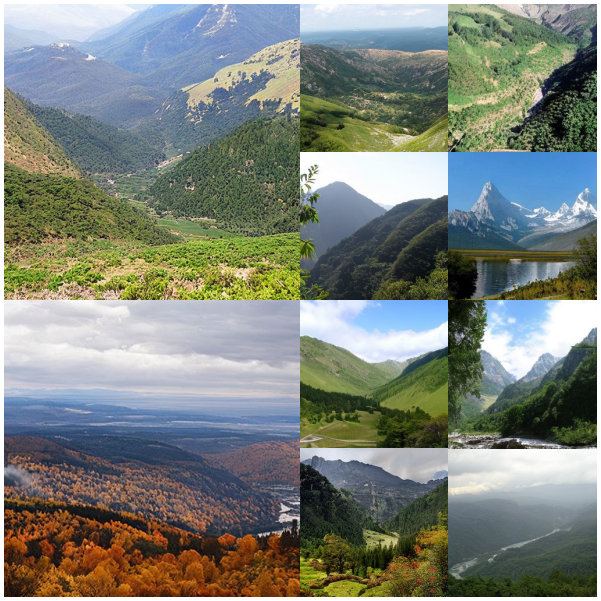

Figure 8. **Generated Samples from xAR.** xAR is able to generate high-fidelity valley (979) images.

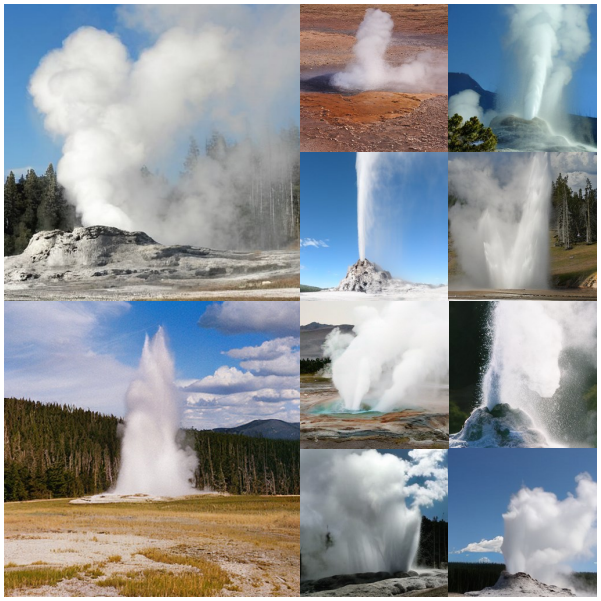

Figure 7. **Generated Samples from xAR.** xAR is able to generate high-fidelity geyser (974) images.

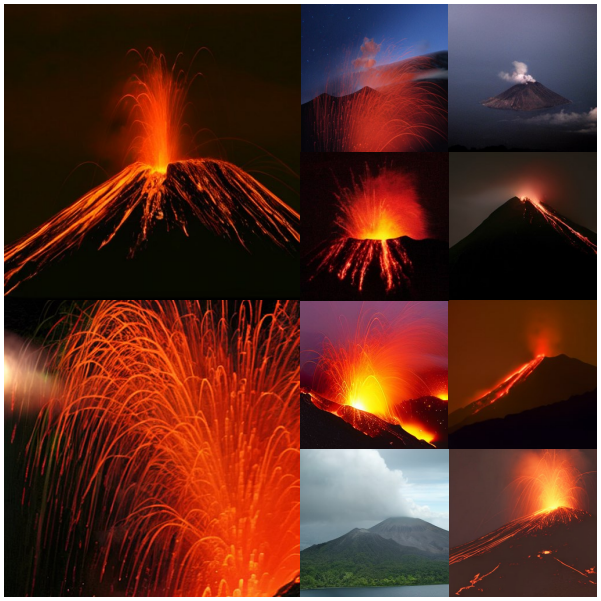

Figure 9. **Generated Samples from xAR.** xAR is able to generate high-fidelity volcano (980) images.

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[2] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *NeurIPS*, 2024. 1

[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. 1

[4] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *ECCV*, 2024. 1

[5] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 1

[6] 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. *arXiv preprint arXiv:2410.06940*, 2024. 1