Neighboring Autoregressive Modeling for Efficient Visual Generation
Supplementary Material

A. Additional Experimental Results

A.1. Speed comparisons with MAR and MaskGIT

A speed comparison between NAR, MAR and MaskGIT is presented in Table A. Each method's default step setting is used.

Table A. Speed comparisons with MAR and MaskGIT. Latency is measured with a batch size of 8.

Model	Steps	Params	Latency
MaskGIT	8	174M	1.36s
MAR-B	64	208M	19.9s
NAR-M	31	290M	**0.33s**

A.2. Benchmarking against LlamaGen with learning rate decay

As shown in Table B, NAR outperforms LlamaGen with the same learning rate scheduler.

Table B. Performance of LlamaGen with learning rate decay.

Model	FID	IS
LlamaGen-B	4.92	206.8
LlamaGen-L	3.31	258.1
NAR-B	4.65	212.3
NAR-L	3.06	263.9

A.3. Comparison with Lformer

While NAR and Lformer [2] share similarities in generation order, they differ fundamentally in the technical design. First, NAR proposes the concept of neighboring autoregressive modeling, which enforces a **strict neighboring constraint**: newly generated tokens have a Manhattan distance of 1 to the tokens generated in the previous step. This constraint is absent in Lformer, which does not explicitly incorporate neighboring relationships. Second, NAR innovates with dimension-oriented decoding heads and mixed logits sampling, which aligns precisely with next-neighbor prediction and enables seamless extension to **video generation**, which is also absent and a non-trivial adaptation for Lformer. Finally, NAR demonstrates superior performance to Lformer, as shown in Table C.

Table C. Performance comparison on MMCelebA-HQ.

Model	Params	FID↓
Lformer-E	1B	18.60
NAR-B	**130M**	**14.66**

B. Discussion on the conditional independence

As noted in [1], conditional independence leads to inconsistent output in parallel decoding. We demonstrate that our proposed mixed logits sampling strategy can mitigate this issue. To illustrate, consider the toy example in Figure 4 of the paper. Let M denote the Transformer backbone, H_h the horizontal head, and H_v the vertical head. The final logits for token $x_{2,1}$ are computed as $\frac{H_h(M(x_{1,1}))+H_v(M(x_{1,0}))}{2}$, while the logits for $x_{2,0}$ are given by $H_h(M(x_{1,0}))$. Assuming $M(x_{1,0})$ follows a multivariate normal distribution, $H_v(M(x_{1,0}))$ and $H_h(M(x_{1,0}))$ are conditionally independent only if $H_v^T H_h = 0$. Note that this condition is overly restrictive and our empirical results show that our trained models do not satisfy this, which justifies the effectiveness of our mixed logits sampling strategy in mitigating conditional independence.

C. More Visualizations

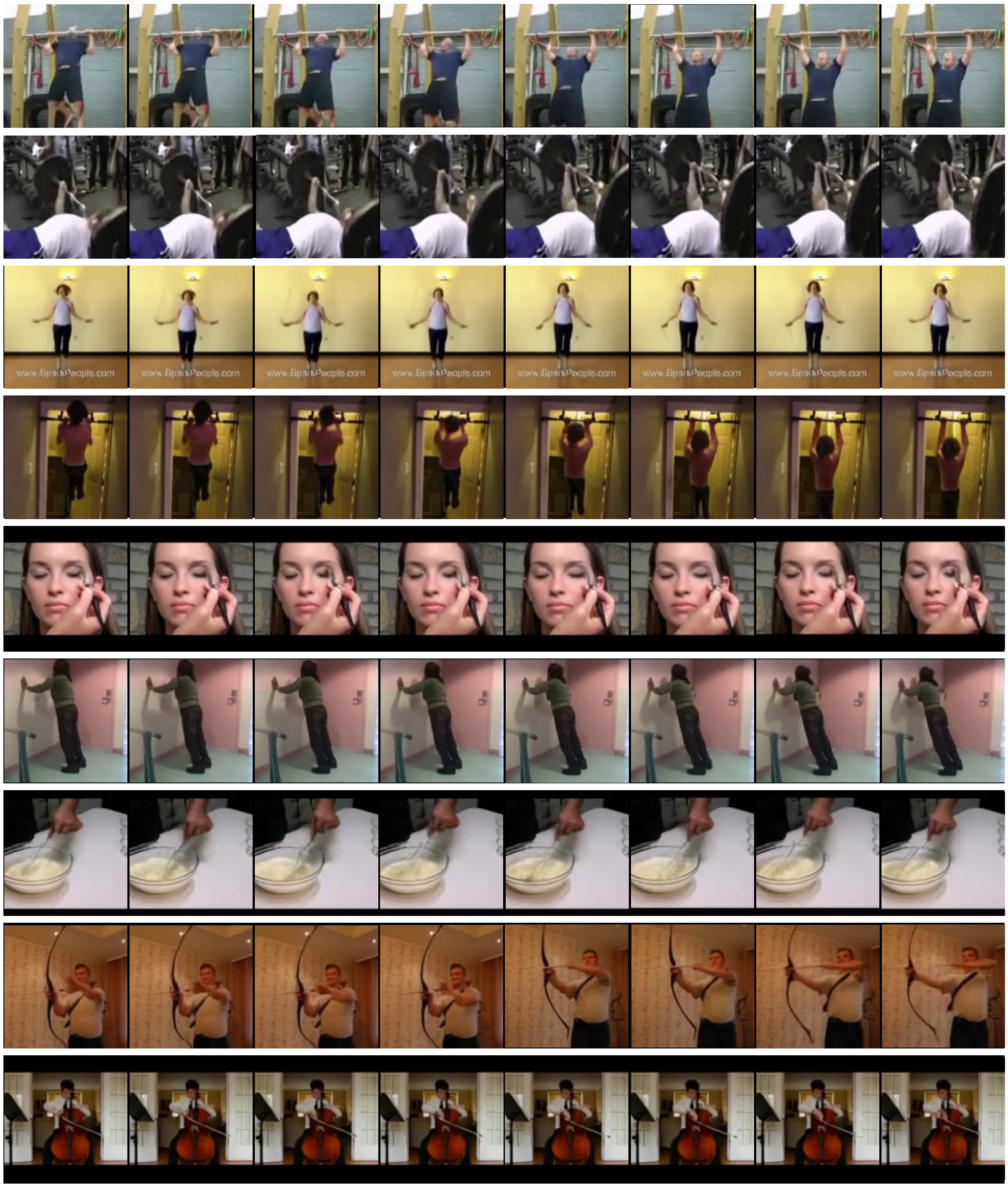

Figure A. **Video generation samples** on UCF-101 dataset. Each row shows sampled frames from a 16-frame, 128×128 resolution sequence generated by NAR-XL across various action categories.

class id 284, siamese cat

class id 973, coral reef

class id 980, volcano

class id 387, lesser panda

class id 979, valley

class id 2, great white shark

class id 985, daisy

class id 974, geyser

Figure B. **Class-conditional image generation samples** produced by NAR-XXL on ImageNet 256×256.

class id 933, cheeseburger

class id 928, ice cream

class id 780, schooner

class id 437, beacon

class id 90, lorikeet

class id 250, Siberian husky

class id 562, fountain

class id 972, cliff

Figure C. **Class-conditional image generation samples** produced by NAR-XXL on ImageNet 256×256.

Figure D. 256×256 **text-guided image generation samples** produced by LlamaGen-XL-Stage1 with next-token prediction paradigm and NAR-XL-Stage1 with next-neighbor prediction paradigm.

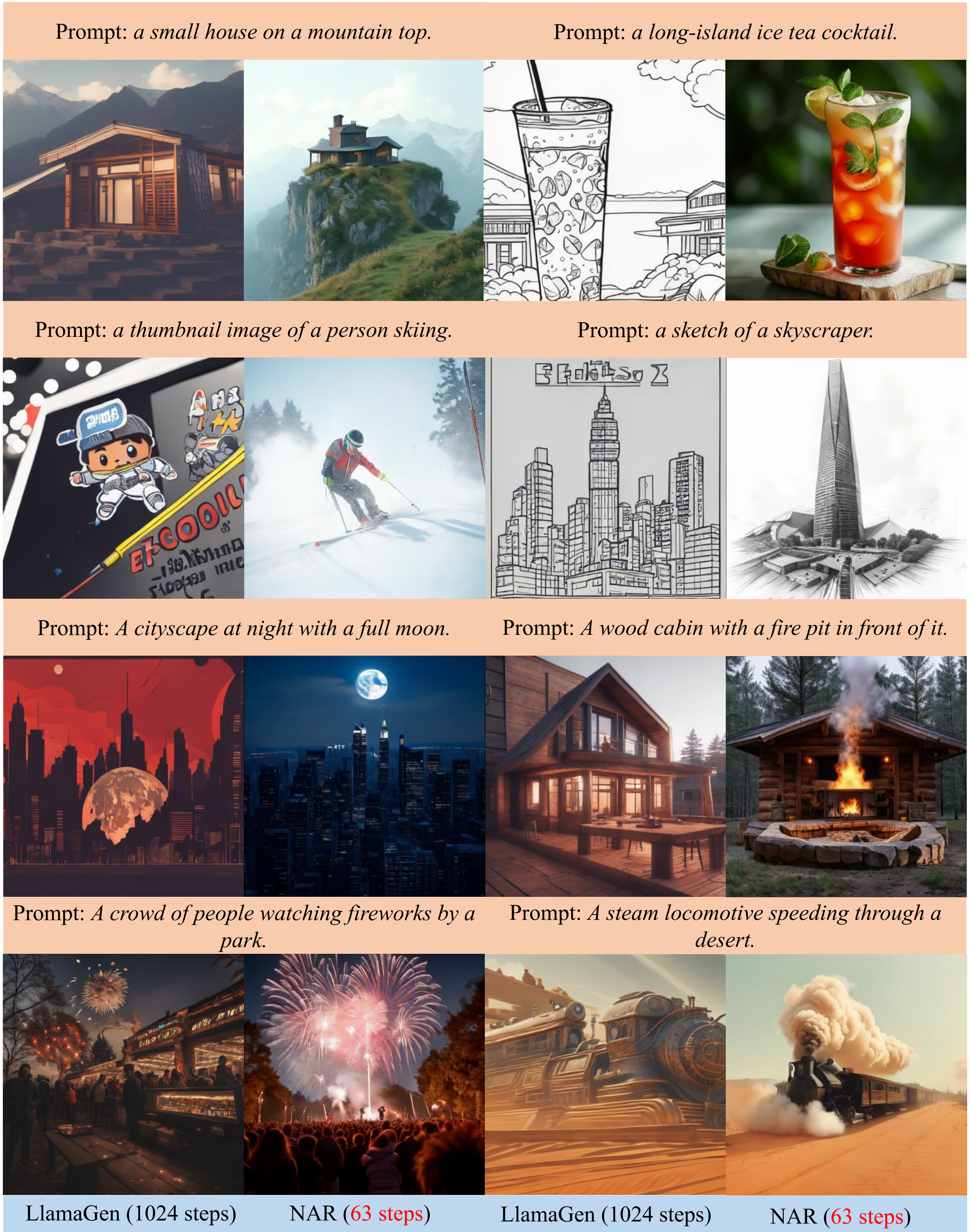

Figure E. 512×512 **text-guided image generation samples** produced by LlamaGen-XL-Stage2 with next-token prediction paradigm and NAR-XL-Stage2 with next-neighbor prediction paradigm. The text prompts are sampled from Parti prompts.

References

[1] Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. Non-autoregressive neural machine translation. *arXiv preprint arXiv:1711.02281*, 2017. 1

[2] Jiacheng Li, Longhui Wei, ZongYuan Zhan, Xin He, Siliang Tang, Qi Tian, and Yueting Zhuang. Lformer: Text-to-image generation with l-shape block parallel decoding. *arXiv preprint arXiv:2303.03800*, 2023. 1