

Collaborative Decoding Makes Visual Auto-Regressive Modeling Efficient

Supplementary Material

In this document, we provide supplementary materials that extend beyond the scope of the main manuscript, constrained by space limitations. These additional materials include:

- We provide more quantitative analysis results to further illustrate our approach;
- We offer more qualitative comparisons for visualization;
- We discuss the limitations of our approach and look into future work.

A. Additional Quantitative Results

In this section, we present additional quantitative analyses to further substantiate our approach.

Impact of Increasing Model Parameters. To validate **Observation 1**, we analyze the effect of varying model sizes on class-conditional image generation quality using ImageNet-256 [1]. Specifically, we evaluate the impact of model size at the k -th scale by predicting the token map r_k with four Visual Autoregressive (VAR) models [7] of different parameter sizes (2B, 1B, 0.6B, and 0.3B). For all other scales ($r_1, r_2, \dots, r_{k-1}, r_{k+1}, \dots, r_{10}$), the largest VAR-d30 model is used for generation. Detailed quantitative results are summarized in Table 1. Our results reveal that increasing model parameters at the earlier scales yields significant improvements in generation quality. However, as the scales progress, the marginal benefits of larger models diminish. At the final scale—responsible for 38% of the sequence tokens—we observe that the performance of the 2B model is nearly identical to that of the 0.3B model. This indicates that as the predicted scale increases, the demand for model parameters to ensure accurate token predictions decreases substantially. These findings highlight significant computational redundancy in the current VAR inference process at larger scales.

Training-Free Performance of CoDe. The proposed CoDe framework employs a large drafter model in conjunction with a smaller refiner model for progressive inference. Notably, it can operate in a training-free manner by leveraging pre-trained VAR-d30 and VAR-d16 models as the drafter and refiner, respectively. Table 3 presents the performance of training-free CoDe across various drafting step settings $N = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Even without additional training, CoDe achieves competitive performance, surpassing the VAR-d24 and VAR-d20 models while maintaining the same speedup ratio.

Image Quality Assessment. In our paper, we use standard metrics such as FID [3], Inception Score (IS), Precision, and Recall to evaluate the generation quality. In or-

Table 1. Impact of increasing parameters across scales

Scale	Params	FID ↓	IS ↑	Precision↑	Recall↑
2	0.3B	2.23	291	0.8122	0.5895
2	0.6B	2.13	292	0.8078	0.5947
2	1.0B	2.04	295	0.8107	0.6027
2	2.0B	1.95	301	0.8107	0.5945
3	0.3B	2.35	283	0.8064	0.5864
3	0.6B	2.21	290	0.8047	0.5967
3	1.0B	2.09	295	0.8074	0.5940
3	2.0B	1.95	301	0.8107	0.5945
4	0.3B	2.27	290	0.8086	0.5953
4	0.6B	2.18	293	0.8068	0.5924
4	1.0B	2.13	296	0.8061	0.5983
4	2.0B	1.95	301	0.8107	0.5945
5	0.3B	2.17	296	0.8119	0.5936
5	0.6B	2.13	298	0.8087	0.5948
5	1.0B	2.10	301	0.8087	0.6025
5	2.0B	1.95	301	0.8107	0.5945
6	0.3B	2.09	301	0.8119	0.5984
6	0.6B	2.05	304	0.8100	0.5976
6	1.0B	2.05	305	0.8089	0.5999
6	2.0B	1.95	301	0.8107	0.5945
7	0.3B	2.09	302	0.8067	0.6010
7	0.6B	2.05	305	0.5095	0.6061
7	1.0B	2.04	307	0.8077	0.6008
7	2.0B	1.95	301	0.8107	0.5945
8	0.3B	2.08	304	0.8135	0.5978
8	0.6B	2.04	308	0.8110	0.6024
8	1.0B	2.02	307	0.8094	0.6038
8	2.0B	1.95	301	0.8107	0.5945
9	0.3B	2.02	304	0.8133	0.6059
9	0.6B	2.01	307	0.8121	0.5948
9	1.0B	2.00	307	0.8097	0.6011
9	2.0B	1.95	301	0.8107	0.5945
10	0.3B	1.99	306	0.8120	0.5978
10	0.6B	1.97	305	0.8102	0.6053
10	1.0B	1.98	303	0.8102	0.6053
10	2.0B	1.95	301	0.8107	0.5945

der to more comprehensively evaluate the quality of generated images, we introduced three image quality assessment (IQA) metrics, including MUSIQ [4], CLIPQA [8], and NIQE [5]. MUSIQ, CLIPQA, and NIQE are three distinct

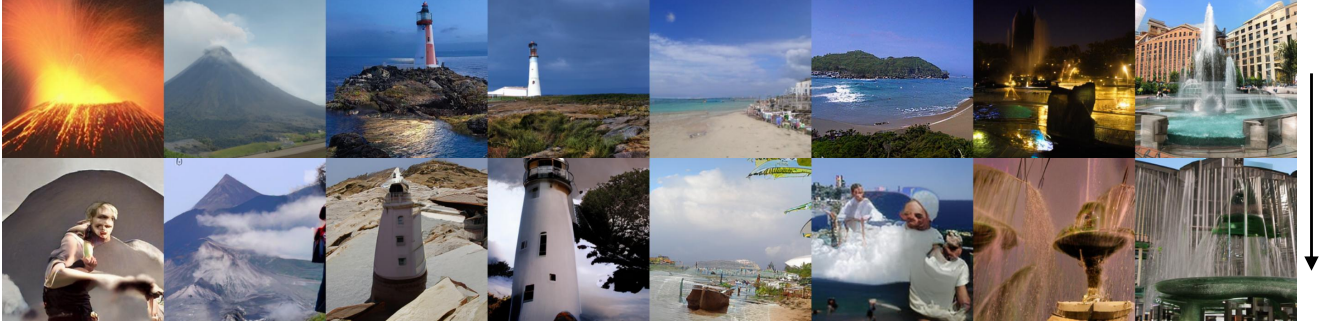


Figure 1. Up: images generated by the original VAR-d16 models. Down: images generated by the perturbation fine-tuned VAR-d16.

Table 2. No reference metrics for additional image quality assessments.

Method	Inference Efficiency						Image Quality Assessment		
	#Steps	Speedup↑	Latency↓	Throughput↑	#Param	Memory↓	MUSIQ ↑	CLIPQA ↑	NIQE↓
VAR-d30	10	1.0x	3.62s	17.71it/s	2.0B	40414MB	60.72	0.6813	6.1739
VAR-CoDe N=9	9+1	1.2x	2.97s	21.54it/s	2.0+0.3B	28803MB	60.78	0.6818	6.1024
VAR-CoDe N=8	8+2	1.7x	2.11s	30.33it/s	2.0+0.3B	21019MB	60.79	0.6812	6.0849
VAR-CoDe N=7	7+3	2.3x	1.60s	40.00it/s	2.0+0.3B	19943MB	60.82	0.6800	6.1247
VAR-CoDe N=6	6+4	2.9x	1.27s	50.39it/s	2.0+0.3B	19943MB	60.76	0.6808	6.1490

Table 3. The training-free performance of CoDe

Configuration	FID ↓	IS ↑	Precision↑	Recall↑
CoDe N=9	1.99	306	0.8120	0.5978
CoDe N=8	2.10	308	0.8155	0.5915
CoDe N=7	2.25	309	0.8204	0.5781
CoDe N=6	2.42	306	0.8283	0.5721
CoDe N=5	2.56	303	0.8313	0.5660
CoDe N=4	2.75	295	0.8342	0.5427
CoDe N=3	2.99	288	0.8410	0.5327
CoDe N=2	3.19	283	0.8433	0.5179
CoDe N=1	3.39	268	0.8132	0.5382

IQA metrics, each with unique approaches and strengths. MUSIQ (Multi-Scale Image Quality) leverages a vision transformer (ViT) [2] and a multi-scale representation to evaluate global aesthetics and local distortions, making it effective for diverse image types, including high-resolution and non-standard aspect ratios. CLIPQA utilizes the pre-trained CLIP [6] model, which combines semantic understanding from large-scale image-text training to assess image quality in a context-aware manner, excelling in tasks aligned with human perception. In contrast, NIQE (Natural Image Quality Evaluator) is a no-reference metric that models natural scene statistics (NSS) using a multivariate Gaussian distribution to measure deviations from high-quality

natural image properties. While MUSIQ and CLIPQA excel in leveraging learned features for state-of-the-art performance, NIQE stands out for its simplicity, computational efficiency, and independence from reference images, though it may struggle with unnatural or heavily edited content. Together, these metrics cater to diverse IQA needs, from deep-learning-based evaluations to lightweight statistical assessments. As shown in Table 2, our CoDe method achieves comparable or even superior generation quality compared to the original VAR-d30. This result further demonstrates the effectiveness of our approach.

B. More Qualitative Results.

Additional Qualitative Comparisons. We provide additional qualitative comparisons between the original VAR-d30 model and our proposed CoDe framework, evaluated with varying drafting steps $N = \{6, 7, 8, 9\}$. As shown in Figures 2 and 3, CoDe achieves significant speedup and substantial memory optimization, with only minimal quality degradation that is nearly imperceptible to the human eye. Even at a speedup rate of 2.9x, the generated images maintain exceptionally high quality and preserve accurate semantic information. It is important to emphasize that the primary goal of CoDe is to enhance the efficiency of the VAR inference process while maintaining high generation quality, rather than reproducing the exact outputs of the original model. Through specialized fine-tuning,

CoDe’s drafter model demonstrates superior predictive accuracy compared to the original model, sometimes resulting in a different global structure. Nevertheless, the image quality remains consistently high and, in some cases, even improves over the original outputs.

Qualitative Results of Perturbation Fine-Tuning. In our study, we conducted a perturbation fine-tuning experiment to examine the distinct generative roles of small and large scales. Using a pre-trained VAR-d16 model, we applied the CSE loss exclusively to tokens in the largest three scales and fine-tuned the model for just 1% of the original training epochs. This minimal fine-tuning at large scales caused a complete collapse of the model’s global modeling capacity at small scales, with the FID increasing from 3.30 to 21.93 and the IS score dropping from 277 to 88. Figure 1 illustrates the qualitative results of perturbation fine-tuning. After slight fine-tuning, the VAR-d16 model nearly loses its ability to model global structures. These findings underscore that VAR models undertake entirely distinct generative tasks at small and large scales, with minimal overlap in functionality.

C. Limitations and Future Work

Limitations. The core concept of CoDe involves decomposing the next-scale prediction process into a collaboration between a large model and a small model. This approach necessitates the availability of two models with different sizes. If only a single large VAR model is available and faster inference is desired, it becomes necessary to retrain a smaller refiner model. However, since the refiner model can be extremely compact, techniques such as model pruning and knowledge distillation can be applied to limit the additional training cost.

Future Work. This study demonstrates that CoDe significantly reduces inference latency and memory consumption for VAR models. Furthermore, the efficiency gains from CoDe become even more pronounced in computationally intensive scenarios. As a result, CoDe is particularly well-suited for high-resolution image generation tasks based on next-scale prediction. In future work, we aim to explore the application of CoDe in building an efficient VAR model specifically optimized for high-resolution image generation.

References

- [1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1
- [2] Kai Han, Yunhe Wang, Hanqing Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunqing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022. 2
- [3] 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. *Advances in neural information processing systems*, 30, 2017. 1
- [4] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 1
- [5] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a “completely blind” image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 1
- [6] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2
- [7] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024. 1
- [8] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 1

Original VAR-d30 1.0x Speedup Throughput: 17.71it/s Memory: 40414MB FID: 1.95



VAR-CoDe N=9 1.2x Speedup Throughput: 21.54it/s Memory: 28803MB FID: 1.94



VAR-CoDe N=8 1.7x Speedup Throughput: 30.33it/s Memory: 21019MB FID: 1.98



VAR-CoDe N=7 2.3x Speedup Throughput: 40.00it/s Memory: 19943MB FID: 2.11



VAR-CoDe N=6 2.9x Speedup Throughput: 50.39it/s Memory: 19943MB FID: 2.27



Figure 2. Qualitative comparison between the original VAR-d30 model and our proposed CoDe model, with different drafting steps.

Original VAR-d30 **1.0x** Speedup Throughput: **17.71it/s** Memory: **40414MB** FID: **1.95**



VAR-CoDe N=9 **1.2x** Speedup Throughput: **21.54it/s** Memory: **28803MB** FID: **1.94**



VAR-CoDe N=8 **1.7x** Speedup Throughput: **30.33it/s** Memory: **21019MB** FID: **1.98**



VAR-CoDe N=7 **2.3x** Speedup Throughput: **40.00it/s** Memory: **19943MB** FID: **2.11**



VAR-CoDe N=6 **2.9x** Speedup Throughput: **50.39it/s** Memory: **19943MB** FID: **2.27**



Figure 3. Qualitative comparison between the original VAR-d30 model and our proposed CoDe model, with different drafting steps.