Depth Adaptive Efficient Visual Autoregressive Modeling

Supplementary Material

A. Ablation on Different Reference Metrics

The reference metric $\mathcal{E}$ and its layer range $[\ell_{\text{begin}}, \ell_{\text{end}}]$ determine the base decision rank map that guides depth allocation. We ablate these choices in Table 6, including metrics analogous to those in SparseVAR [5] ($\mathcal{E}_{\text{MSE}}$ on Block 3) and FastVAR [24] ($\mathcal{E}_{\text{SUB}}$). While ($\mathcal{E}_{\text{MAE}}$, $[3, 19]$) and ($\mathcal{E}_{\text{MSE}}$, $[0, 31]$) are mostly comparable, the MAE metric offers a more balanced trade-off between quality and latency. This highlights that the effectiveness of our framework stems from how ranks are utilized for dynamic depth scheduling, rather than the precision of the initial rank estimation itself.

Table 6. Ablation study on reference metrics. We compare different metrics, including those analogous to SparseVAR ($\mathcal{E}_{\text{MSE}}$, Block 3) and FastVAR ($\mathcal{E}_{\text{SUB}}$), under different reference scales $r_{\mathcal{R}}$.

$\mathcal{R}$	Reference Metric		GenEval		ImageReward	
	$\mathcal{E}$	$[\ell_{\text{begin}}, \ell_{\text{end}}]$	Score↑	Avg Latency (ms)↓	Score↑	Avg Latency (ms)↓
7	$\mathcal{E}_{\text{MAE}}$	$[3, 19]$	0.7256	1168	0.9088	1174
	$\mathcal{E}_{\text{MAE}}$	$[0, 31]$	0.7216	1228	0.9081	1253
	$\mathcal{E}_{\text{MSE}}$	$[3, 19]$	0.7219	1217	0.9094	1214
	$\mathcal{E}_{\text{MSE}}$	$[0, 31]$	0.7304	1270	0.8948	1295
	$\mathcal{E}_{\text{MSE}}$	Block 3	0.7198	1164	0.9078	1184
	$\mathcal{E}_{\text{SUB}}$	–	0.7210	1242	0.9033	1230
8	$\mathcal{E}_{\text{MAE}}$	$[3, 19]$	0.7262	1295	0.9171	1303
	$\mathcal{E}_{\text{MAE}}$	$[0, 31]$	0.7215	1339	0.9148	1353
	$\mathcal{E}_{\text{MSE}}$	$[3, 19]$	0.7241	1336	0.9172	1348
	$\mathcal{E}_{\text{MSE}}$	$[0, 31]$	0.7300	1377	0.9108	1384
	$\mathcal{E}_{\text{MSE}}$	Block 3	0.7226	1302	0.9155	1307
	$\mathcal{E}_{\text{SUB}}$	–	0.7207	1371	0.9152	1392
9	$\mathcal{E}_{\text{MAE}}$	$[3, 19]$	0.7318	1622	0.9254	1616
	$\mathcal{E}_{\text{MAE}}$	$[0, 31]$	0.7198	1651	0.9292	1670
	$\mathcal{E}_{\text{MSE}}$	$[3, 19]$	0.7282	1660	0.9236	1663
	$\mathcal{E}_{\text{MSE}}$	$[0, 31]$	0.7310	1686	0.9231	1692
	$\mathcal{E}_{\text{MSE}}$	Block 3	0.7276	1623	0.9271	1638
	$\mathcal{E}_{\text{SUB}}$	–	0.7323	1663	0.9261	1689

B. Sensitivity Analysis of Reference Ranges

We conduct a sensitivity analysis to assess how the choice of the reference layer range $[\ell_{\text{begin}}, \ell_{\text{end}}]$ impacts generation quality. To map the optimization landscape, we compute the mean and standard deviation of SSIM scores on 100 prompts generated by expanding class names [32, 41], for both MAE and MSE metrics, with a range step of 2. As illustrated in Figs. 9 and 10, the response patterns are consistent across different computational constraints ($\mathcal{R} = 7, 8$), indicating that the optimal layer range is independent of the overall compute budget. While MSE yields a marginally higher peak SSIM, MAE demonstrates a broader performance plateau, resulting in a slightly better average SSIM across all configurations (Fig. 11a).

To validate these observations on benchmarks, we select four representative configurations on MAE as highlighted in Fig. 11b: (A) our default setting $[3, 19]$, (B) a single-layer reference $[8, 8]$ yielding the highest mean SSIM, (C) a statistical optimum $[10, 16]$ balancing mean and std, and (D) a sub-optimal region $[26, 28]$. Table 7 reports their performance on GenEval [20] and ImageReward [63]. Notably, while the statistically optimal trade-off (C) yields the highest scores, sufficient baseline performance is maintained as long as the configuration resides within the stable plateau (setting A) and avoids poor choices like low-performing regions (setting D) or an insufficient number of reference lay-

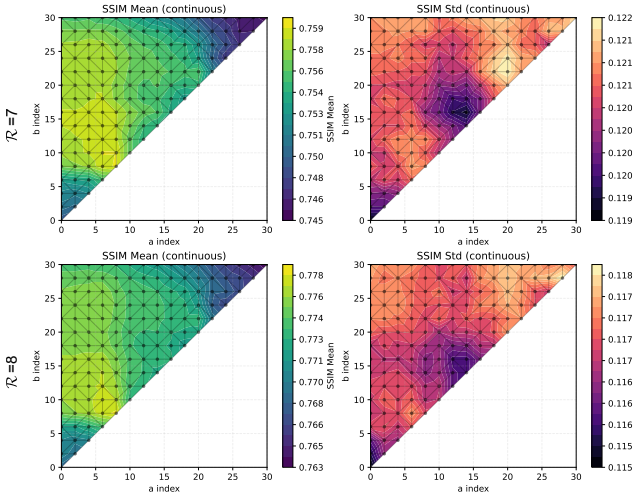


Figure 9. Sensitivity analysis of $[\ell_{\text{begin}}, \ell_{\text{end}}] = [a, b], \mathcal{E} = \mathcal{E}_{\text{MAE}}$.

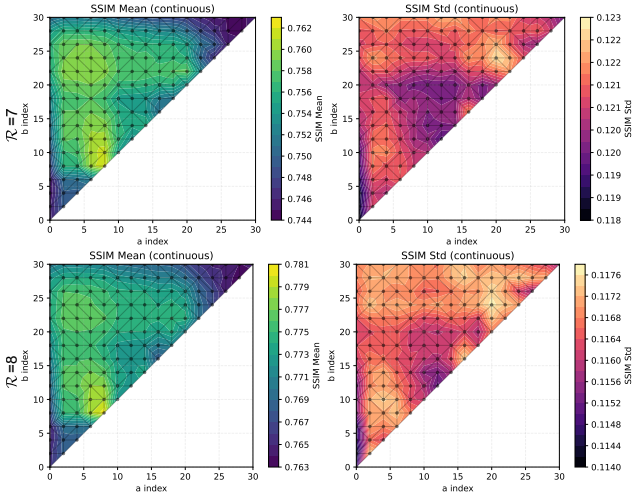
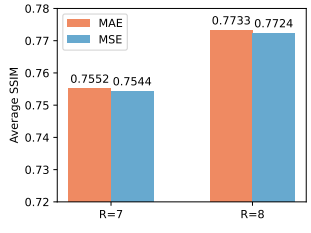
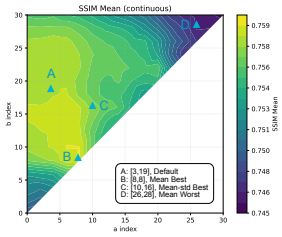


Figure 10. Sensitivity analysis of $[\ell_{\text{begin}}, \ell_{\text{end}}] = [a, b], \mathcal{E} = \mathcal{E}_{\text{MSE}}$.



(a) Comparison of mean SSIM averaged over ranges.



(b) Selection of four layer range configurations.

Figure 11. (a) Comparison of reference metrics, where $\mathcal{E}_{\text{MAE}}$ shows a marginally better average SSIM. (b) Selection of four specific layer configurations: (A) our default, (B) best SSIM mean, (C) best SSIM mean+std, and (D) near-worst SSIM.

Table 7. Performance of selected layer range configurations on GenEval and ImageReward benchmarks. The configurations A, B, C, and D correspond to the points highlighted in Fig. 11b.

$\mathcal{R}$	Selection	$[\ell_{\text{begin}}, \ell_{\text{end}}]$	GenEval Overall ↑	Image Reward ↑
7	A	[3, 19]	0.7256	0.9088
	B	[8, 8]	0.7179	0.9096
	C	[10, 16]	0.7283	0.9101
	D	[26, 28]	0.7220	0.9016
8	A	[3, 19]	0.7262	0.9171
	B	[8, 8]	0.7260	0.9179
	C	[10, 16]	0.7313	0.9188
	D	[26, 28]	0.7259	0.9095

Table 8. Performance comparison of bit-reversal and uniform configurations on GenEval, ImageReward, and HPSv2.1 benchmarks.

$\mathcal{R}$	Configuration	GenEval		ImageReward		HPSv2.1	
		Overall ↑	Latency (ms) ↓	Overall ↑	Latency (ms) ↓	Score ↑	Latency (ms) ↓
7	bit-reversal	0.7256	1168	0.9088	1174	30.06	1185
	uniform	0.7255	1207	0.9070	1218	30.01	1232
8	bit-reversal	0.7262	1295	0.9171	1303	30.16	1285
	uniform	0.7258	1346	0.9156	1356	30.13	1340
9	bit-reversal	0.7318	1622	0.9254	1616	30.29	1625
	uniform	0.7263	1677	0.9291	1670	30.27	1667

ers (setting B). This observation, together with the SSIM sensitivity visualization (Fig. 9), highlights DepthVAR's low sensitivity to its hyperparameters, ensuring that suboptimal choices do not lead to catastrophic performance degradation.

C. Ablation on Bit-reversal

We validate the design of our layer selection mechanism by comparing our bit-reversal permutation against a uniform sampling strategy. The uniform baseline distributes active layers equidistantly, whereas our bit-reversal method functions as a quasi-Monte Carlo sampler. As shown in Table 8, the bit-reversal configuration consistently outperforms the uniform selection across all benchmarks and reference scale with slightly lower latencies, albeit with small margins. These results confirm the benefits of distributing active layers with bit-reversal.

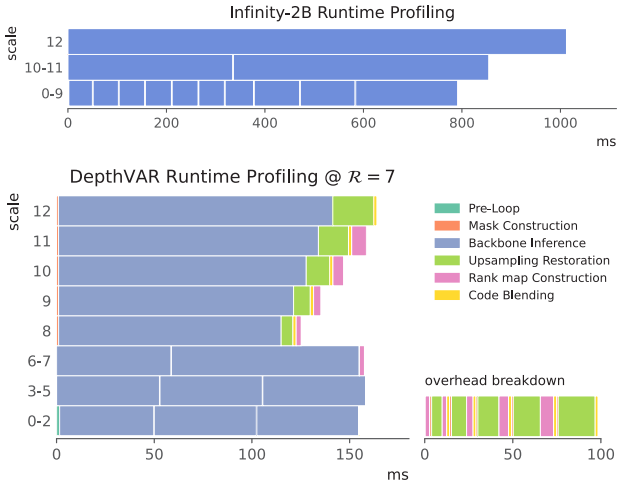
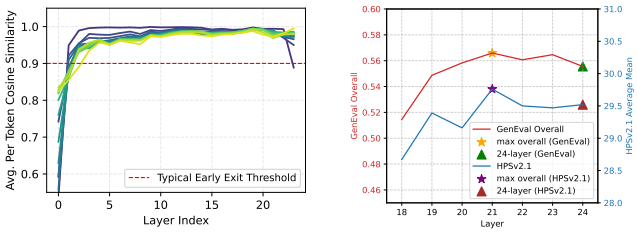


Figure 12. Runtime breakdown analysis on Infinity [25].

D. More Efficiency Profiling

As illustrated in Fig. 12, we profile the runtime overhead of DepthVAR on Infinity with $\mathcal{R}$=7. The breakdown shows that our acceleration framework introduces ∼100ms of additional overhead, dominated by rank-map construction and upsampling operations, which represent roughly 6% of the total computation savings. Furthermore, caching intermediate layer behaviors from the previous scale incurs approximately 1.1GB of peak GPU memory overhead. The overall memory footprint is reduced **from 16.5GB to 11.6GB**, achieving 4.9GB savings and outperforming FastVAR [24]'s 4.2GB reduction by an additional 700MB.



(a) Token-wise Layer Similarity.

(b) Early-exiting Evaluations.

Figure 13. Generalization of depth redundancy to the HART [54] architecture. (a) Token-wise representation similarity exhibits saturation patterns consistent with standard VAR models. (b) Early-exiting evaluations on GenEval [20] and HPSv2.1 [60] confirm that generation quality peaks prior to the final layer.

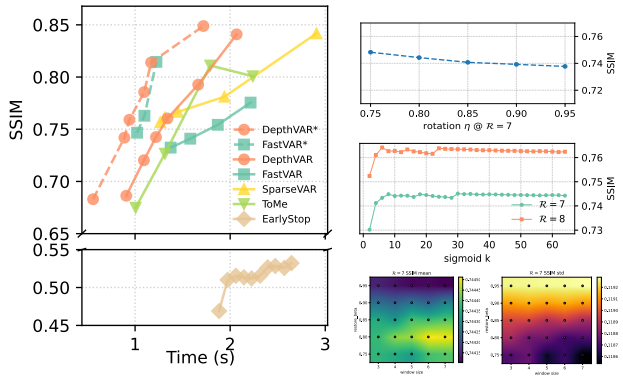


Figure 14. SSIM–latency pareto frontier. Dashed *: w/o last 2 scales. DepthVAR consistently traces the upper-left.

Figure 15. Sensitivity analysis over additional hyperparameters.

(a) Finer Detail Loss.

(b) Dense Structure Handling.

Figure 16. Qualitative failure cases. (a) Loss of fine-grained details. (b) Difficulty with complex dense structures.

E. Universality of Depth Redundancy

To show that the observed depth redundancy is a fundamental property of VAR models, we extend our analysis from Sec. 3.1 to the HART [54] architecture, a distinct hybrid variant. Applying the same evaluation protocols, we conducted token-wise layer similarity and early-exiting analyses. As shown in Fig. 13, the results confirm that HART exhibits similar redundancy patterns, with generation quality peaking at 21 before the full layer. This suggests that depth redundancy is a pervasive characteristic of visual autoregressive models, validating the broader applicability of our adaptive depth paradigm to other architectures.

F. Extended Pareto Frontier Analysis

While individual points provide specific quality-speed results, our method's superiority is best characterized by the Pareto frontier, as in Fig.14. We analyze the overall structural fidelity of DepthVAR on the Infinity [25] following the evaluation protocol in Appx.B. An observation of simple early-stopping truncation is that it significantly alters the generative trajectory, making it not a direct substitute if aiming to preserve structural identity. To ensure a fair comparison with FastVAR [24], we report both full-scale and omitting the final two scales. In both cases, DepthVAR traces the upper-left envelope, demonstrating the trade-off.

G. More Hyperparameter Sensitivity

We provide additional analysis for the remaining hyperparameters in Fig. 15, showing their properties. We observe that the rotation magnitude η primarily balances overall structural quality against semantic alignment. For the schedule sharpness k, SSIM exhibits local peaks near 8, 18, and 30, making $k \in [8, 30]$ a robust range for stable inference. Finally, minimal sensitivity to the restoration window size and β is observed, suggesting potential simplification.

H. More Qualitative Visualizations

The qualitative visualizations in the main text (Fig. 6) are derived from the ImageReward [63] benchmark. We provide additional results from HPSv2.1 [60] in Fig. 17. These examples further demonstrate that DepthVAR preserves high visual fidelity and rich detail, reinforcing its superior speed-quality trade-off compared to hard-pruning methods. Despite its performance, DepthVAR has limitations. As shown in Fig. 16, it can struggle with fine-grained details and dense structures, likely because its fixed compute budget is insufficient for universally complex images. This limitation points toward future work in dynamic compute allocation, as discussed in our conclusion.

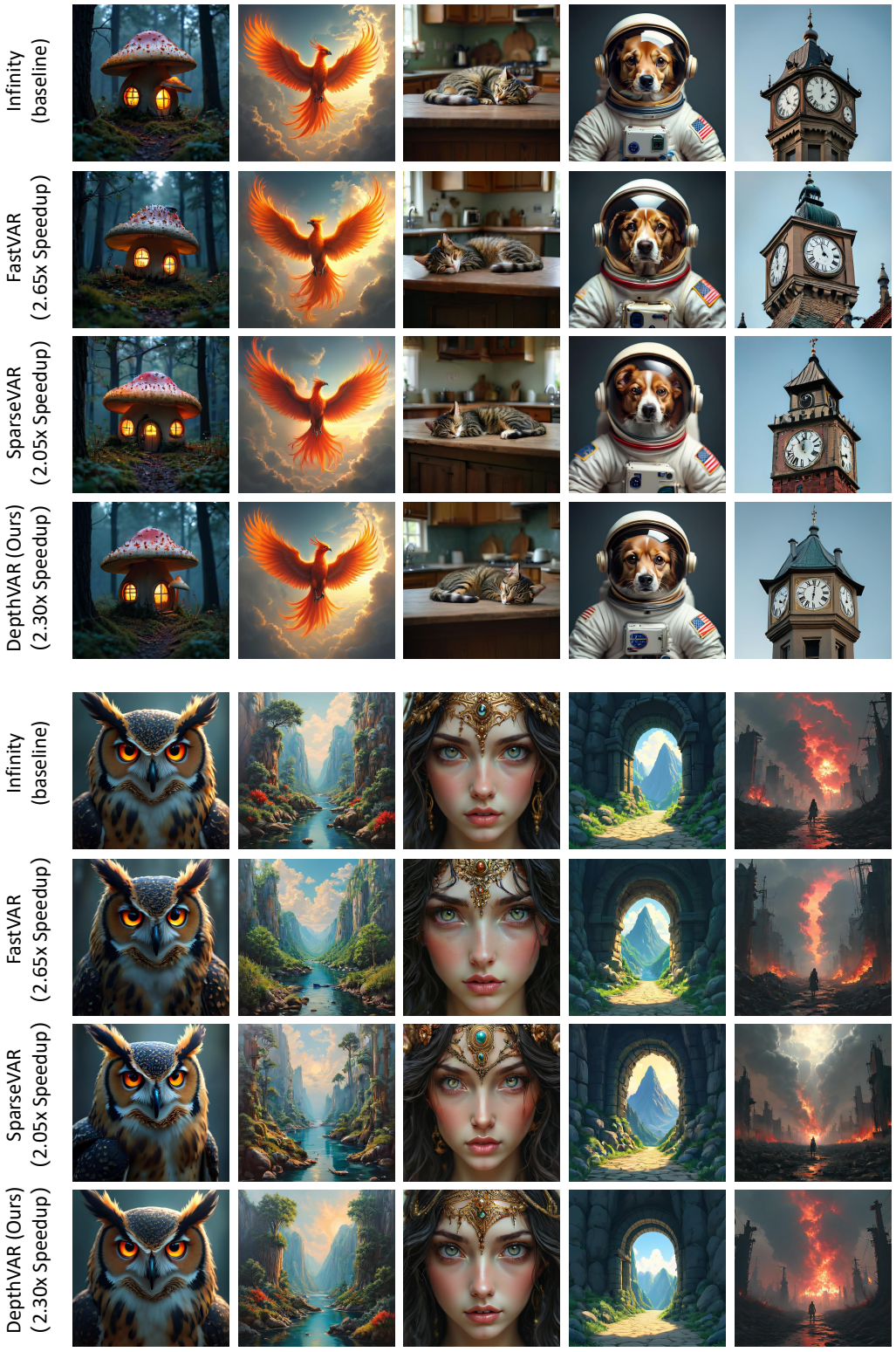


Figure 17. Additional qualitative visual comparisons from the HPSv2.1 benchmark. DepthVAR consistently preserves visual fidelity and semantic details, demonstrating a superior speed-quality trade-off.