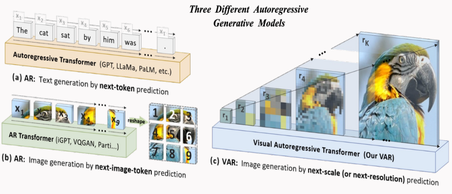
ToProVAR: Efficient Visual Autoregressive Modeling via Tri-Dimensional Entropy-Aware Pruning

Jiayu Chen, Ruoyu Lin, Jingxin Li, Xiang Chen

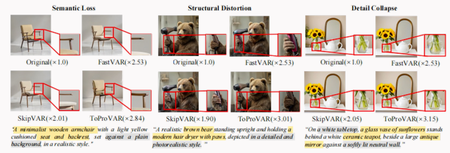
School of Computer Science, Peking University

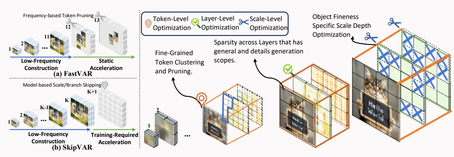
Background



- Traditional AR: **token-by-token**, strictly sequential → **slow & non-parallelizable**.
- VAR: generate **coarse→fine multi-scale tokens** → faster, but **cost grows exponentially** with scale.

Problem





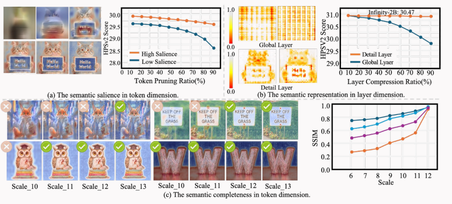
Limitations of 1D Optimizations (FastVAR&SkipVAR)

- Only optimize one dimension (token / scale).
- Cannot capture **semantic relations** → structure distortion.
- Remove tokens that need **deeper scale** → detail collapse.

3D Sparsity Optimizations (ToProVAR)

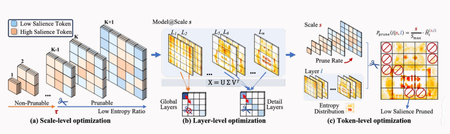
- We propose a **multi-dimensional optimization framework** that analyzes and prunes computation across **token, layer, and scale** dimensions.

Analysis



- **Token-level:** Few tokens carry key semantics; pruning **90%** low-salience tokens yields **<2%** quality loss.
- **Layer-level:** Global layers are pruning-sensitive, while detail layers can be pruned by **up to 90%**.
- **Scale-level:** Objects require different scale depths for correct semantic modeling.

Method



Scale-level optimization via Semantic Fineness.

As scales grow, semantics stabilize.

Measure stability using the **low-entropy ratio**:

$$\rho_s = \frac{|\{i \mid H_i^s < \overline{H}^s\}|}{N_s}.$$

Pruning starts once

$$D = \min\{s \mid \rho_s \geq \tau\}.$$

→ Only scales deeper than D are pruned.

Layer-level optimization via Semantic Scope.

Use SVD to distinguish **Global Layers** vs. **Detail Layers**:

$$X = U\Sigma V^\top, \quad \varrho^{(l,s)} = \frac{\sigma_1}{\sigma_2}.$$

Define representation score:

$$\mathcal{R}^{(l,s)} = \exp\big(-\beta(\varrho^{(l,s)} - 1)\big).$$

→ Prune only layers with **high** $\mathcal{R}^{(l,s)}$ (Detail Layers).

Token-level optimization via Fine-grained Semantic Salience.

Normalize token entropy:

$$\hat{H}_i^{(l,s)} = \frac{H_i^{(l,s)}}{\sum_j H_j^{(l,s)}}.$$

Compute unified pruning tendency:

$$q_i^{(l,s)} = \phi(s) \cdot \mathcal{R}^{(l,s)} \cdot \hat{H}_i^{(l,s)}, \quad \phi(s) = \frac{s}{S_{\max}}.$$

Retention probability:

$$P_{\text{keep}}(i \mid s, l) = \begin{cases} 1, & s < D, \\ 1 - \text{clip}(\alpha_{\min} + (\alpha_{\max} - \alpha_{\min})q_i, 0, 1), & s \geq D. \end{cases}$$

→ High-entropy tokens in deep scales & detail layers are pruned first.

Main Result

Quantitative comparison on GenEval and DPG.

Methods	GenEval					DPG				Latency(s)↓	Speedup
	Two Obj.	Position	Color Attri.	Overall↑	Entity	Relation	Attribute	Overall↑		
Infinity-2B	79.01	24.00	58.00	0.69	90.81	88.19	87.89	83.41	2.10	1.0 ×
+FastVAR	78.79	27.75	59.50	0.68	88.86	**91.57**	87.46	83.39	0.80	2.6 ×
+SkipVAR	76.77	26.50	57.50	0.67	**89.30**	87.07	87.01	82.94	1.10	2.0 ×
+ToProVAR	**75.80**	**29.50**	**62.00**	**0.69**	87.39	88.92	**90.01**	**83.07**	**0.61**	**3.4** ×
Infinity-8B	96.97	61.00	75.00	0.83	90.92	93.57	88.83	86.68	4.86	1.0 ×
+FastVAR	94.19	57.00	75.25	0.81	90.80	**92.30**	90.40	86.50	2.01	2.4 ×
+SkipVAR	94.94	57.50	**76.50**	0.82	89.71	90.52	90.02	86.44	2.11	2.3 ×
+ToProVAR	**94.95**	**61.00**	76.00	**0.83**	**91.11**	90.39	**91.04**	**86.70**	**1.78**	**2.7** ×

Quantitative Comparison on HPSv2 and ImageReward.

Methods	HPSv2.1					ImageReward↑	Latency(s)↓	Speedup
	Photo	Concept-Art	Anime	Paintings	Overall↑			
Infinity-2B	29.40	30.38	31.72	30.39	30.47	0.94	1.57	1.0 ×
+FastVAR	28.86	29.90	31.12	29.92	29.95	0.92	0.62	2.5 ×
+SkipVAR	29.25	**30.25**	31.50	**30.45**	**30.39**	0.93	0.87	1.8 ×
+ToProVAR	**29.26**	30.15	**31.44**	30.23	30.27	**0.93**	**0.58**	**2.7** ×
Infinity-8B	29.42	31.27	32.45	30.83	30.99	1.04	5.31	1.0 ×
+FastVAR	**29.87**	30.42	31.80	29.89	30.24	1.02	1.97	2.6 ×
+SkipVAR	29.09	30.86	32.04	**30.55**	**30.64**	1.03	2.65	2.0 ×
+ToProVAR	29.19	**30.89**	**32.09**	30.24	30.58	**1.04**	**1.75**	**3.0** ×

Quantitative comparisons of FID and CLIP score on the MJHQ30K dataset.

Method	Landscape			People			Food		
	Latency	FID↓	CLIP↑	Latency	FID↓	CLIP↑	Latency	FID↓	CLIP↑
Infinity-2B	1.67	44.1	0.267	1.71	58.91	0.281	1.69	84.2	0.270
+FastVAR	0.60	45.1	0.264	0.61	71.8	0.274	0.61	84.7	0.273
+SkipVAR	1.01	58.1	0.260	1.06	73.7	0.253	0.88	102.3	0.256
+ToProVAR	**0.50**	**44.5**	**0.264**	**0.48**	**58.84**	**0.283**	**0.46**	**84.3**	**0.274**

Ablation study of the Three-Dimensional Progressive Manipulation Framework on Infinity-2B.

Method	Latency(s)↓	Speed ↑	GenEval↑
Infinity-2B	2.10	-	0.690
+ Scale Depth Loc.	0.47	4.5 ×	0.477
++ Layer Repr. Ident.	0.57	3.7 ×	0.679
+++ Fine-grained Token Prun.	0.61	3.4 ×	0.690

Visualization

Prompt:a dog holds a board with the text 'ToProVAR is Efficient'

Reference

[1] Zhuokun Chen, Jugang Fan, Zhuowei Yu, Bohan Zhuang, and Mingkui Tan. Frequency-aware autoregressive modeling for efficient high-resolution image synthesis, 2025a.

[2] Hang Guo, Yawei Li, Taolin Zhang, Jiangshan Wang, Tao Dai, Shu-Tao Xia, and Luca Benini. Fastvar: Linear visual autoregressive modeling via cached token pruning.

[3] Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024.

[4] Jiajun Li, Yue Ma, Xinyu Zhang, Qingyan Wei, Songhua Liu, and Linfeng Zhang. Skipvar: Accelerating visual autoregressive modeling via adaptive frequency-aware skipping. arXiv preprint arXiv:2506.08908, 2025a.

[5] Ziran Qin, Youru Lv, Mingbao Lin, Zeren Zhang, Danping Zou, and Weiyao Lin. Head-aware kv cache compression for efficient visual autoregressive modeling, 2025.