ExtraVAR: Stage-Aware RoPE Remapping for Resolution Extrapolation in Visual Autoregressive Models

Feihong Yan[1] Shaoyu Liu[2] Haixuan Wang[3] Shuai Lu[1,6]
Weijie Ma[1] Linfeng Zhang[4] Huiqi Li[1,6,*] Xiangyang Ji[5,*]

[1] Beijing Institute of Technology [2] Xidian University [3] Northeastern University at Qinhuangdao
[4] Shanghai Jiao Tong University [5] Department of Automation, Tsinghua University
[6] Beijing Key Laboratory of Intelligent Diagnosis Technology Equipment for Optic Nerve-Related Eye Diseases, Beijing, China

* Corresponding authors

NeurIPS 2026

Three failures under resolution extrapolation

We change one band of the rotary position embedding (RoPE) at a time.

	Global repetition	Local repetition	Detail degradation
Control			
Intervention			
	Layout construction Low-frequency rotation disabled	**Local refinement** Mid-band wavelengths set to one-sixth of the native token-grid side length	**Detail refinement** High-frequency rotation disabled

Frequency bands across generation stages

Layout construction



Global layout and color

Local refinement



Object composition

Detail refinement



Edges and texture

Intermediate outputs; later panels show enlarged crops.

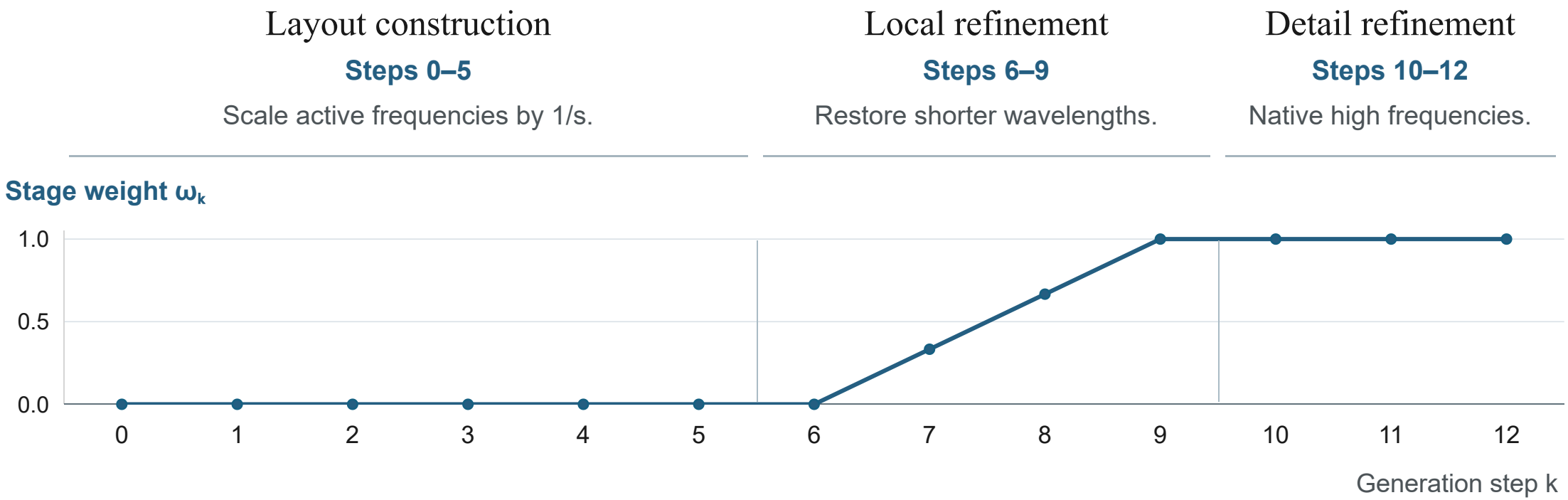
	Scale id 1	2	3	4	5	6	7	8	9	10	11	12
Very-Low (supportive)	**0.31**	**0.30**	0.27	0.26	0.23	0.21	0.21	0.20	0.18	0.16	0.15	0.15
Low (global)	0.29	**0.39**	**0.44**	**0.43**	**0.41**	**0.37**	**0.34**	**0.31**	0.27	0.24	0.22	0.21
Mid (local)	0.21	0.18	0.18	0.21	0.25	**0.30**	**0.32**	**0.34**	**0.35**	**0.33**	**0.32**	**0.30**
High (detail)	0.19	0.13	0.11	0.10	0.11	0.12	0.13	0.15	0.20	0.27	**0.31**	**0.34**

Relative query ℓ_2 norm

Columns show steps 1–12.
At the three marked steps,
the largest relative norms occur
in the low, mid and high bands,
respectively.

Stage-aware frequency remapping

The stage weight controls how far high frequencies recover their native values.



Low, mid and high bands $\theta_{k,j} = \theta_j \left[s^{-1} + \omega_k \rho_j (1 - s^{-1}) \right]$

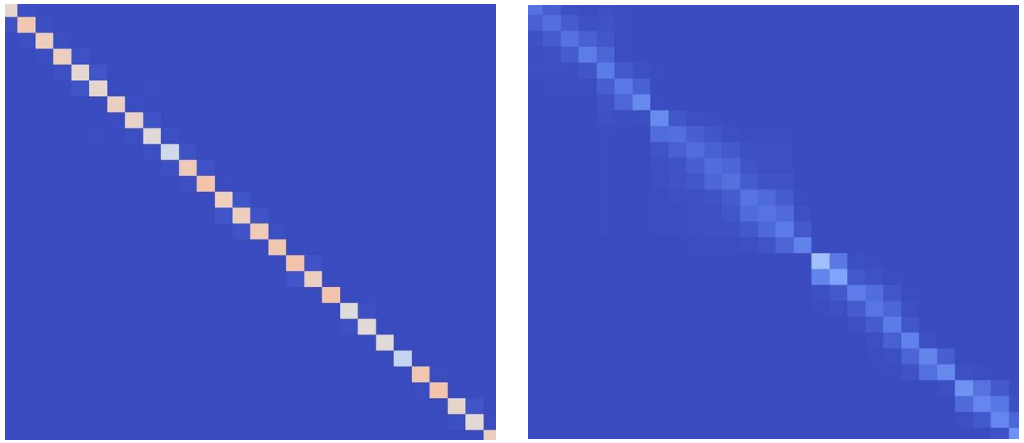
s = L′/L is the target-to-native token-grid side ratio.
θ_j is the native frequency of rotary pair j.
ρ_j ranges from 0 in the low band to 1 in the high band.

Very-low-frequency pairs retain their query and key features with zero rotation throughout.

Attention calibration during detail refinement

We compare each head's concentration at the native and target resolutions.

Effect of a high-band intervention



Control

High-band rotation disabled

First-order logit scaling

Divide each head's mean row entropy by the log of its key count.

$$\hat{\alpha} = 1 + [\mathcal{H}(1) - \mathcal{H}^{\mathrm{tr}}] \log N_k / V_g$$

H(1) and H^{tr} are target and native normalized entropies.
V_g is the attention-weighted logit variance, averaged over queries.

Scale logits by $\hat{\alpha}$ when $k > 9$, $H^{tr} < 0.3$ and $V_g \geq \varepsilon$.
Otherwise, use $\alpha = 1$.

ε is a small threshold for numerical stability.

Object structure at two target resolutions

	Peng et al. (2024)	Issachar et al. (2025)	**ExtraVAR**
Nesting dolls 2048 × 2048			
Blueberry tart 2560 × 2560			

Each row uses the same prompt and frozen Infinity-2B checkpoint across methods.

Benchmark performance on Infinity-2B

Infinity-2B: native 1024 × 1024, evaluated at 2560 × 2560.
All methods use the same frozen checkpoint, prompts, sampler and cache settings.

Metric ↑	Best prior score	ExtraVAR	Gain
GenEval	0.65	**0.70**	+0.05
DPG-Bench	82.69	**83.58**	+0.89
HPSv2.1	28.40	**30.11**	+1.71

Best prior is the highest of the five same-backbone baseline scores for each metric. Gains are absolute differences.

Component ablation and transfer

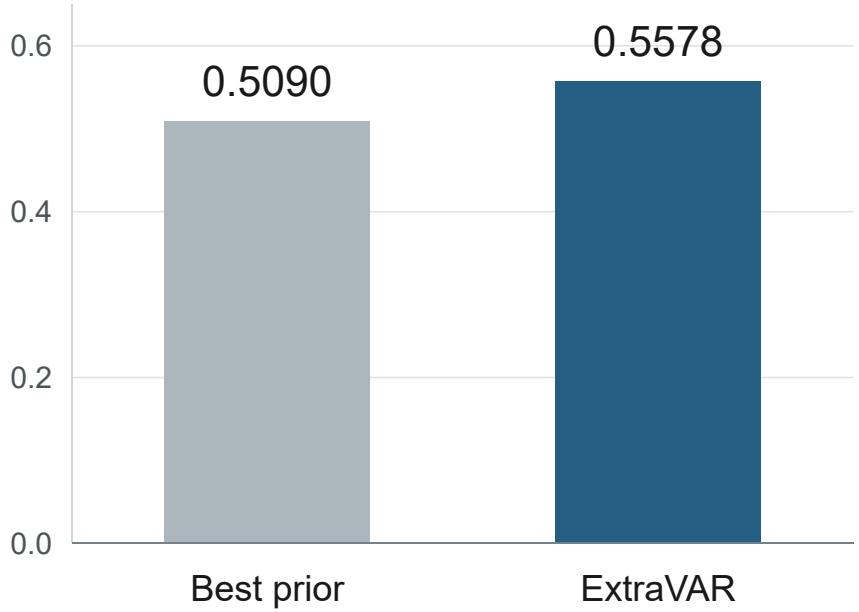
GenEval at the same target resolution on two backbones.

Infinity-2B component ablation

Frequency map	Calibration	Color attr.	Overall
Uniform scaling	Off	0.40	0.57
Uniform scaling	On	0.42	0.58
Stage-aware	Off	**0.53**	0.69
Stage-aware	**On**	**0.59**	**0.70**

Calibration improves color attribution from 0.53 to 0.59 after remapping.

Transfer to Switti-1024



Switti uses the same band definitions and its own stage boundaries.

Conclusion

As the image takes shape, the dominant RoPE band shifts from low through mid to high frequencies.

We stretch active wavelengths during layout construction, then restore shorter wavelengths and calibrate attention during detail refinement.

github.com/feihongyan1/ExtraVAR