DVAR: Dynamic Visual Autoregressive Modeling for Image Super-Resolution

Supplementary Material

In this Supplementary Material, we provide comprehensive details to further validate the effectiveness of our approach. First, Section A details the user study settings and results. Section B provides a thorough complexity analysis comparing our DVAR with state-of-the-art methods. In Section C, we present additional experimental results focusing on the ablation of dynamic components. We discuss the limitations of the proposed method in Section D. Finally, extensive visual comparisons are provided in Section E.

A. User Study

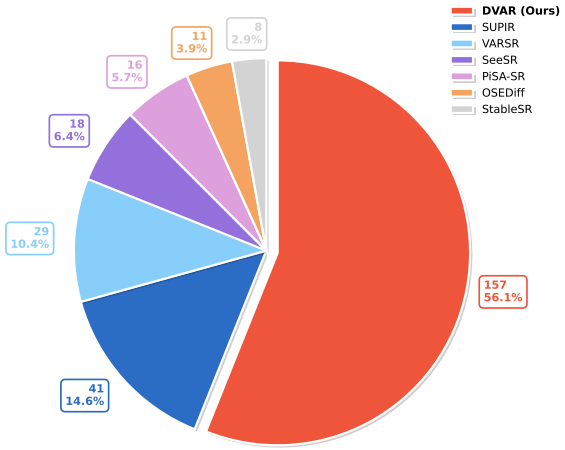

Figure 6. **User Study Results.** The distribution of first-choice selections by participants comparing DVAR with six other state-of-the-art methods.

To comprehensively evaluate the perceptual quality of our method, we conducted a user study. We randomly selected 20 real-world images from the RealSR [4] and DRealSR [40] datasets. We compared our DVAR with six state-of-the-art methods, including diffusion-based approaches (StableSR [37], SUPIR [31], SeeSR [42], OSEDiff [41], PiSA-SR [31]) and the visual autoregressive-based method (VARSR [27]). We invited 14 volunteers to participate in the study. For each test case, participants were presented with the original LQ image alongside the seven restored results, displayed in a randomized order to prevent positional bias. They were asked to select the single best result based on overall perceptual realism and fidelity to the original content. In total, we collected and analyzed 280 votes (14 participants $\times$ 20 images).As shown in Figure 6, DVAR achieved the highest selection rate of 56.1%, significantly outperforming all competing methods. This decisive preference indicates that the images generated by DVAR exhibit a superior balance of perceptual richness and content fidelity, aligning most closely with human visual preference in real-world super-resolution scenarios.

B. Complexity Analysis

Table 4. Complexity comparison of diffusion- and VAR-based methods.

Method	Inf. Step	Inf. Time (s)	Param (M)
StableSR [37]	200	16.72	1409
DiffBIR [23]	50	8.45	1717
SeeSR [42]	50	6.36	2284
SUPIR [46]	50	12.87	4801
PiSA-SR [31]	1	0.27	1775
VARSR [27]	10	0.45	1102
DVAR(Ours)	10	1.13	2682

We analyze the computational complexity of DVAR in terms of inference speed and model parameters, with results presented in Table 4. As a pure visual autoregressive framework, DVAR demonstrates substantial speed advantages over iterative diffusion models, achieving approximately 5 to 12$\times$ acceleration compared to methods like SeeSR and SUPIR, while maintaining competitive speed with single-step methods, trailing them by less than one second. Regarding model size, DVAR's larger parameter count is a direct reflection of its powerful backbone, Switti [35], a large-scale autoregressive foundation model. Despite the increased capacity, our approach offers a distinct advantage in training efficiency and accessibility over VARSR. Unlike VARSR, which necessitates full fine-tuning and the retraining of its VQ-VAE decoder on large-scale private datasets, DVAR utilizes parameter-efficient Low-Rank Adaptation and operates on pre-trained components without modification. This non-invasive design drastically reduces training costs and ensures full reproducibility on public data.

C. Additional Experimental Results

We also provide visual comparisons of the DVAR variants in Figure 8. Consistent with the quantitative results, all DVAR variants recover sharper local textures than VARSR on standard 512$\times$512 inputs (*e.g.*, the details in the eye region). However, variants without explicit size conditioning exhibit noticeable structural distortions when applied to other target resolutions. By contrast, the full DVAR preserves both faithful details and structural coherence across

Table 5. Additional multi-resolution results on RealSR

Resolution	SeeSR					PiSA-SR					DVAR(Ours)				
	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MANIQA↑	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MANIQA↑	PSNR↑	SSIM↑	LPIPS↓	CLIPIQA↑	MANIQA↑
256×256	21.06	0.5347	0.4489	0.3645	0.3739	**22.82**	**0.5930**	0.2931	0.6162	0.4566	21.56	0.5848	**0.2806**	**0.6936**	**0.6224**
384×384	**24.43**	**0.6929**	**0.2703**	0.6401	0.5190	23.77	0.6507	0.2881	0.6143	0.4780	22.69	0.6428	0.3061	**0.7005**	**0.6057**
512×288	24.86	0.7088	0.3034	0.6190	0.5024	**25.64**	**0.7427**	**0.2671**	0.6608	0.4800	23.25	0.6541	0.3574	**0.7056**	**0.5952**
512×384	25.13	0.7199	0.2963	0.6309	0.5083	**25.51**	**0.7417**	**0.2671**	0.6657	0.4799	23.45	0.6733	0.3343	**0.6978**	**0.5836**
512×512	25.21	0.7221	0.3000	0.6676	0.5402	**25.50**	**0.7418**	**0.2672**	0.6697	0.4810	23.71	0.6811	0.3291	**0.7050**	**0.5756**

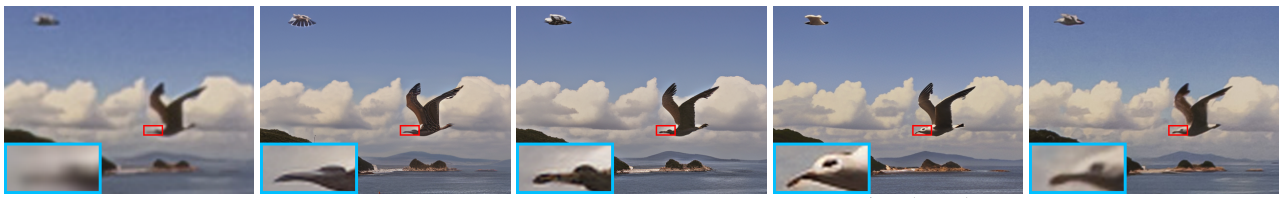

Figure 7. Ablation visual results of the aspect ratio 4:3.

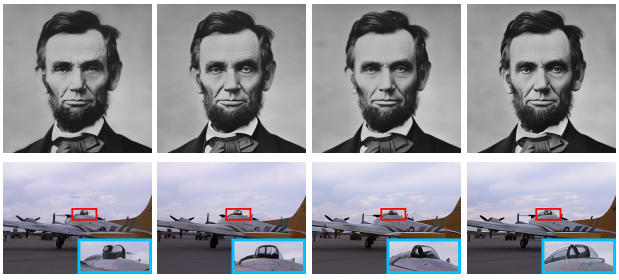

Figure 8. Ablation results of the Dynamic Components.

all tested scales. These observations further validate the necessity of size conditioning for robust size-flexible inference across variable input sizes.

To further complement the ablation study in the main paper, we additionally report a *single-model* cross-resolution evaluation on paired RealSR over multiple sizes and aspect ratios (Tab. 5). Compared with representative diffusion-based Real-ISR baselines, DVAR exhibits markedly stronger robustness to changes in target resolution. In particular, when the target resolution changes from 512×512 to 256×256, the average degradation of DVAR is only **2.8%** on full-reference metrics and shows no average degradation on no-reference metrics (**-3.3%**), compared with 30.7%/38.1% for SeeSR and 13.4%/6.5% for PiSA-SR. Representative qualitative results at 4:3 (512×384) are shown 7.

D. Limitations

Despite its strong performance, DVAR has several limitations. First, the quality of the derived text prompt is contingent upon the degradation-robust encoder's output. In instances of severe degradation, the encoder can produce a flawed reconstruction, yielding an erroneous prompt that may, in turn, induce semantic deviations in the final high-resolution output. Second, as a purely autoregressive model, DVAR operates within the well-documented perception-distortion trade-off, prioritizing perceptual realism over pixel-wise fidelity metrics such as PSNR and SSIM. While this leads to lower scores on the latter, the model's correspondingly strong performance on LPIPS indicates that perceptual fidelity is well-preserved. Third, consistent with other large-scale generative models, DVAR can struggle to faithfully reconstruct fine-grained scene text. The robust rendering of textual details from low-resolution inputs remains a significant open research challenge for this entire class of models. Finally, while our canonical scaling dynamic provides a strong inductive bias for size generalization, we find that the purely transformer-based backbone still requires explicit size conditioning and training on multi-resolution data to achieve optimal performance across a wide range of sizes. Future work could explore architectural modifications that more natively integrate this scaling logic, thus reducing the reliance on explicit conditioning.

E. More Visualization Results

In Figure 9, we provide extensive qualitative comparisons to further demonstrate the robustness of DVAR. Across a variety of challenging scenarios, our method consistently restores fine textures, structural details, and natural color tones with superior visual fidelity. In contrast, competing methods frequently suffer from blurry outputs, distorted edges, or over-smoothed regions. These comparisons vividly showcase the generalization ability of DVAR in handling complex degradations while preserving intricate image details.

Figure 9. Additional qualitative comparisons with different SOTA methods (Part 2). **Zoom in for a better view.**