AlignVAR: Towards Globally Consistent Visual Autoregression for Image Super-Resolution

Supplementary Material

This supplementary material provides additional analyses and results that extend the findings in the main paper. To clarify the procedural details of our method, §Sec. 1 presents the complete training and inference pipelines of AlignVAR. To assess computational efficiency, §Sec. 2 provides both theoretical and empirical complexity analyses. To further verify perceptual quality, §Sec. 3 reports the results of a user study comparing AlignVAR with several state-of-the-art baselines. To evaluate real-world robustness, §Sec. 4 includes additional experiments on the challenging RealLR200 dataset. To better understand the behavior of our model, §Sec. 5 visualizes multi-scale attention maps produced by AlignVAR. To examine the contribution of each proposed component, §Sec. 6 conducts an ablation analysis of SCA and HCC using the Edge IoU metric. To investigate how SCA affects the receptive range, §Sec. 7 compares attention maps with and without SCA. To analyze the influence of the hyperparameter λ, §Sec. 8 studies its effect on loss convergence and prediction accuracy across scales. To further showcase visual improvements, §Sec. 9 presents extended qualitative comparisons. Finally, §Sec. 10 discusses the limitations of AlignVAR and outlines possible directions for future work.

1. Training and Inference Procedures

This section provides the procedural details of AlignVAR during both training and inference. All notations follow those introduced in the main paper. The overall training pipeline is summarized in Algorithm 1, and the autoregressive reconstruction process is presented in Algorithm 2.

Training procedure. During training, AlignVAR follows a scale-wise teacher-forcing strategy, as outlined in Algorithm 1. Given an HR–LR pair, the frozen VAE encoder [9] extracts the full-resolution latent $z \in \mathbb{R}^{C\times H_K\times W_K}$, where C denotes the channel dimension and (H_K, W_K) is the resolution of the last scale. For each scale k, the corresponding ground-truth latent $u_k^{\text{gt}} \in \{1,\dots,|V|\}^{H_k\times W_k}$ is obtained via spatial downsampling and quantization, where (H_k, W_k) denote the height and width of scale k, and $|V|$ denotes the size of the codebook. The residual token is defined as $r_{\text{gt}}^k = u_{\text{gt}}^k - u_{\text{gt}}^{k-1}$. The autoregressive predictor p_θ receives the reweighted context $\tilde{r}_{\text{gt}}^{1:k-1}$, where the modulation mask $m_k \in \mathbb{R}^{1\times H_k\times W_k}$ is produced by the mask generator M_ϕ using the structural guidance extracted from the LR input. The model predicts the categorical distribution over the $|V|$ codebook entries for r_k, and the parameters of p_θ and M_ϕ are jointly optimized using the multi-scale cross-entropy loss $\mathcal{L}_{\text{CE}}$ and the hierarchical consistency loss $\mathcal{L}_{\text{HCC}}$ applied to the cumulative predictions.

Algorithm 1 AlignVAR Training

Inputs: HR–LR image pair $(I_{\text{HR}}, I_{\text{LR}})$
Hyperparams: steps K, resolutions $\{(H_k, W_k)\}_{k=1}^K$
$z = \mathcal{E}(I_{\text{HR}})$; $c = \mathcal{E}_{con}(I_{\text{LR}})$; $s = |\text{Laplacian}(I_{\text{LR}})|$;
$u_{\text{gt}}^0 = 0$, $u_{\text{pred}}^0 = 0$;
for $k = 1,\dots,K$ **do**
  $u_{\text{gt}}^k = \text{Quant}(\text{Down}(z, H_k, W_k))$;
  $r_{\text{gt}}^k = u_{\text{gt}}^k - u_{\text{gt}}^{k-1}$;
end for
for $k = 1,\dots,K$ **do**
  $\bar{s}_k = \text{norm}(\text{Down}_k(s))$;
  $m_k = \sigma(M_\phi([\, r_{\text{gt}}^k,\ \bar{s}_k\,]))$;
  $\tilde{r}_{\text{gt}}^k = (1 + m_k) \odot r_{\text{gt}}^k$;
  Predict $p_\theta(\tilde{r}_k \mid \tilde{r}_{\text{gt}}^{1:k-1}, c)$;
  $u_{\text{pred}}^k = u_{\text{pred}}^{k-1} + \hat{r}_{\text{pred}}^k$;
end for
Compute $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{HCC}}$;
Update (θ, ϕ);
return trained model parameters;

Inference procedure. During inference, the model performs fully autoregressive multi-scale prediction, as detailed in Algorithm 2. The LR input is first encoded into the conditional latent $c \in \mathbb{R}^{C\times H_K\times W_K}$, and the structural guidance map is computed in the same way as during training. Starting from $u_{\text{pred}}^0 = 0$, the model sequentially predicts the residual tokens $\hat{r}_k \in \{1,\dots,|V|\}^{H_k\times W_k}$, generates the mask m_k, computes the reweighted token map $\tilde{r}_k$, and updates the cumulative latent state u_{pred}^k. After all scales are completed, the final latent is mapped through the shared VQ codebook and decoded by the VAE decoder to produce the super-resolved output.

2. Complexity Analysis

Theoretical complexity. We analyze the computational cost of AlignVAR by examining its multi-scale autoregressive generation process. Let the latent resolution sequence be $\{(h_1, w_1), (h_2, w_2), \dots, (h_K, w_K)\}$, where (h_k, w_k) denotes the height and width of the VQ code map

Algorithm 2 AlignVAR Inference

Inputs: LR image I_{LR}
Hyperparams: steps K, resolutions $\{(H_k, W_k)\}_{k=1}^{K}$
$c = \mathcal{E}_{con}(I_{\text{LR}})$; $s = |\text{Laplacian}(I_{\text{LR}})|$;
$u_{\text{pred}}^{0} = 0$;
for $k = 1, \ldots, K$ **do**
  $\bar{s}_k = \text{norm}(\text{Down}_k(s))$;
  $\hat{r}_k = \arg\max p_\theta(\tilde{r}_k \mid \tilde{r}_{1:k-1}, c)$;
  $m_k = \sigma(M_\phi([\hat{r}_k, \bar{s}_k]))$;
  $\tilde{r}_k = (1 + m_k) \odot \hat{r}_k$;
  $u_{\text{pred}}^{k} = u_{\text{pred}}^{k-1} + \hat{r}_k$;
end for
$f_K = \text{lookup}(V, u_{\text{pred}}^{K})$;
$I_{\text{SR}} = D(f_K)$;
return reconstructed SR image I_{SR};

at the k-th autoregressive step, and the final resolution satisfies $h_K = h$, $w_K = w$. For simplicity, we assume $n_k = h_k = w_k$ for all scales. Following the progressive-resolution design [8], we set $n_k = a^{k-1}$ with a constant $a > 1$, chosen such that $a^{K-1} = n$.

At scale k, the model attends to all accumulated token maps $(r_1, r_2, \ldots, r_k)$, and the total number of tokens is

$$\sum_{i=1}^{k} n_i^2 = \sum_{i=1}^{k} a^{2(i-1)} = \frac{a^{2k} - 1}{a^2 - 1}. \tag{1}$$

Thus, the computational cost of the k-th autoregressive step is

$$\left(\frac{a^{2k} - 1}{a^2 - 1} \right)^2. \tag{2}$$

Summing over all steps yields the total generation complexity:

$$\begin{aligned} &\sum_{k=1}^{\log_a(n)+1} \left(\frac{a^{2k} - 1}{a^2 - 1} \right)^2 \\ &= \frac{(a^4 - 1)\log n + (a^8 n^4 - 2a^6 n^2 - 2a^4(n^2 - 1) + 2a^2 - 1)\log a}{(a^2 - 1)^3 (a^2 + 1)\log a} \\ &\sim \mathcal{O}(n^4). \end{aligned} \tag{3}$$

Since the final resolution satisfies $a^{K-1} = n$, the total complexity is dominated by the last autoregressive step, yielding an overall complexity of $O(n^4)$.

Next, we consider the additional components introduced by AlignVAR, including the structural-guidance mask generator M_ϕ and the residual modulation operation. The mask map at scale k is denoted by $m_k \in \mathbb{R}^{1 \times n_k \times n_k}$. Computing m_k requires a single forward pass over the feature map, followed by element-wise modulation, both of which scale linearly with n_k^2. Since this is asymptotically smaller than the autoregressive cost $\left(\sum_{i=1}^{k} n_i^2 \right)^2$, the additional modules do not influence the overall complexity.

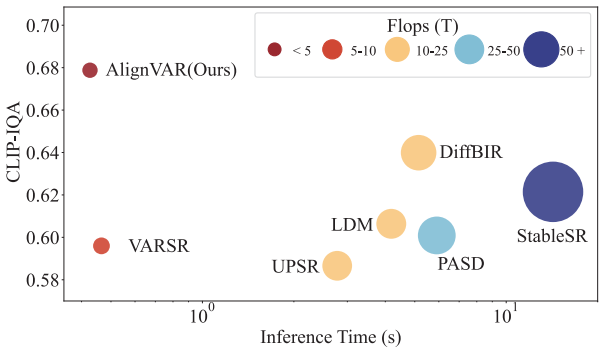

Figure 1. Computational comparison among diffusion-based SR methods (DiffBIR [4], LDM [7], PASD [17], UPSR [20]), the VAR-based baseline (VARSR [6]), and our AlignVAR. AlignVAR achieves high performance with substantially lower complexity than other models.

In summary, the theoretical generation complexity of AlignVAR remains $O(n^4)$, showing that the proposed consistency-enhancing modules maintain the same asymptotic computational cost.

Empirical comparison. To further evaluate the practical complexity, we compare AlignVAR with representative diffusion-based SR methods and the VAR-based baseline VARSR. We measure inference time, FLOPs, and perceptual quality (CLIP-IQA [10]) under the same hardware and input configuration. As shown in Fig. 1, AlignVAR achieves the best accuracy–efficiency balance: it delivers significantly higher CLIP-IQA than diffusion models while maintaining much lower FLOPs and substantially faster inference. Compared with VARSR, AlignVAR attains a large perceptual gain while introducing negligible computational overhead, consistent with our theoretical analysis. These results demonstrate that AlignVAR retains the computational advantages of VAR while notably enhancing global consistency and reconstruction quality.

3. User Study

To evaluate the perceptual quality of the super-resolution results generated by AlignVAR compared to other methods, we conducted a user study. We randomly selected 50 images from three different datasets: RealSR [12], DRealSR [14], and DIV2K-val [1]. These images were then presented to 20 participants, who were asked to compare the super-resolution results produced by AlignVAR against those from several competing models. The models evaluated in the study included VARSR [6], PASD [17], DiffBIR [4], UPSR [20], StableSR [11], and LDM [7]. Participants were asked to choose which model's output they preferred for each image. The results of the user study are shown in Fig. 2, which highlight the superiority of

Table 1. Comparison with state-of-the-art methods on RealLR200. The best and second-best results are highlighted in **bold red** and underline blue, respectively.

Datasets	Metrics	GAN-based			Diffusion-based					VAR-based	
		BSRGAN	Real-ESR	SwinIR	LDM	StableSR	DiffBIR	PASD	UPSR	VARSR	AlignVAR
RealLR200	NIQE↓	4.3817	4.2048	4.2157	4.2533	4.2516	4.1715	4.9330	4.7606	4.4579	**4.0617**
	MANIQA↑	0.5462	0.5582	0.3741	0.3049	0.5841	0.6066	0.5902	0.4206	0.4536	**0.6237**
	CLIPIQA↑	0.5679	0.5389	0.5596	0.5253	0.6068	**0.6797**	0.6509	0.6397	0.6144	0.6734
	MUSIQ↑	64.87	62.94	63.55	55.19	63.30	68.20	62.06	66.46	62.12	**69.36**

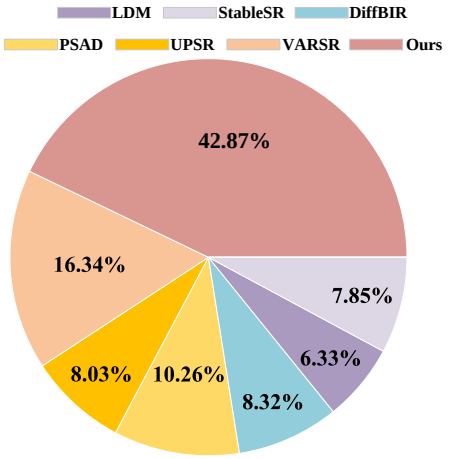

Figure 2. User study results comparing the super-resolution outputs of different models.

AlignVAR in producing high-quality reconstructed images as judged by human evaluators, further validating the effectiveness of the proposed method.

4. Evaluation on the RealLR200 Dataset

To further assess the robustness and real-world applicability of AlignVAR, we conduct additional experiments on the RealLR200 [15] dataset, a challenging collection of real-world low-resolution images with diverse and complex degradations. We compare AlignVAR against representative GAN-based methods (BSRGAN [19], Real-ESR [13], SwinIR [3]), diffusion-based approaches (LDM [7], StableSR [11], DiffBIR [4], PASD [17], UPSR [20]), and the VAR-based baseline VARSR [6]. Both quantitative and qualitative comparisons are performed.

Table 1 summarizes the no-reference quality evaluation across four widely used perceptual metrics: NIQE [5], MANIQA [16], CLIP-IQA [10], and MUSIQ [2]. AlignVAR consistently achieves the best or second-best performance among all competing methods. Notably, AlignVAR surpasses the diffusion-based models, which typically excel in perceptual realism, demonstrating a superior balance between fidelity, sharpness, and naturalness. Compared with VARSR [6], AlignVAR obtains substantial improvements across all metrics, highlighting the effectiveness of the proposed consistency-enhancing mechanisms.

Qualitative results on RealLR200 are presented in Fig. 7. GAN-based models introduce unnatural textures or hallucinated structures, while diffusion-based methods may produce inconsistent details under strong degradations. In contrast, AlignVAR reconstructs coherent edges, stable textures, and visually pleasing details. The improvements are evident in object boundaries and repeated patterns, where spatial and hierarchical consistency play a crucial role.

5. Attention Map Visualization

To further demonstrate the effectiveness of our method in expanding the attention range and enhancing the reweighting process, we visualize the attention maps at various scales. Specifically, we display the attention maps for all scales except for the first one, as shown in Fig. 3. From the visualizations, it is evident that our approach does not limit the attention to the diagonal elements but instead extends the attention across the entire spatial domain at all scales. This highlights the ability of our model to effectively reweight and capture dependencies over larger regions as the resolution increases. The expansive attention observed at higher scales is a clear indication that the reweighting mechanism in AlignVAR facilitates a more global understanding of the image structure.

6. Ablation Study on SCA and HCC

We perform an ablation study to assess the individual and combined impact of SCA and HCC on the performance of AlignVAR. Specifically, we evaluate the Edge Intersection over Union (Edge IoU) [18] across different scales, computed between the predicted edge map and the ground-truth edge map. The edge maps are extracted using the Canny edge detector, and the IoU is then calculated on the resulting binary edge masks to quantify the accuracy of structural recovery. Fig. 4 presents the results of the ablation study. The full AlignVAR model, which incorporates both SCA and HCC, consistently achieves the highest Edge IoU. As the scale increases, the benefit becomes more apparent, demonstrating the importance of enforcing multi-scale con-

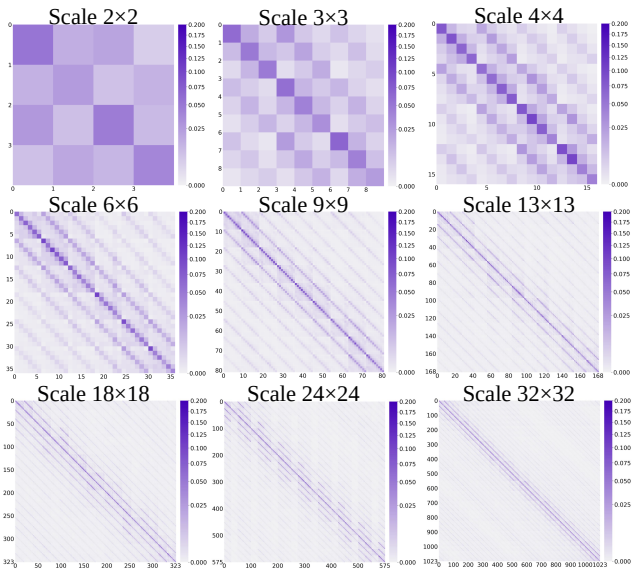

Figure 3. Attention maps at different scales. AlignVAR expands the attention range, capturing dependencies over large areas.

sistency in high-resolution reconstruction. When comparing the variants, removing SCA leads to the most significant performance drop. This indicates that SCA plays a primary role in expanding the spatial attention range and maintaining spatial consistency, which is essential for recovering fine-grained and structurally aligned edges. Removing HCC also results in a decline in Edge IoU, though the degradation is less severe. This suggests that HCC mainly contributes to stabilizing cross-scale dependencies, but its influence on edge localization is comparatively weaker than that of SCA.

7. Impact of SCA on Attention Range

To assess the influence of the SCA on the attention range, we conduct an ablation study by comparing the attention maps with and without the inclusion of SCA. The results are presented in Fig. 5. Without SCA, the attention maps at different scales demonstrate a more localized focus, with attention primarily concentrated along the diagonal. This suggests that, in the absence of SCA, the model tends to focus on nearby regions, limiting its ability to capture long-range dependencies across the spatial domain. In contrast, when SCA is incorporated, the attention maps exhibit a noticeable expansion. The model begins to attend to a much broader area beyond the diagonal, with significant attention given to regions that are further apart. This comparison demonstrates that SCA significantly increases the attention range, allowing the model to better capture both local and global dependencies, which contributes to the improved performance of AlignVAR.

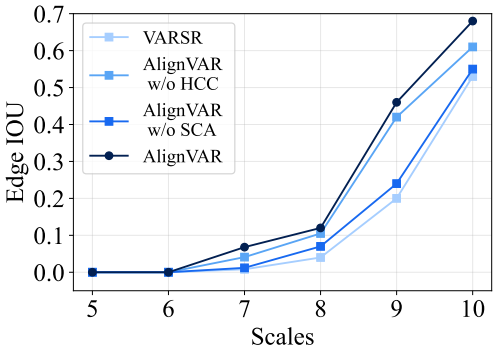

Figure 4. Ablation results of SCA and HCC evaluated using Edge IoU [18] across different scales. The full AlignVAR achieves the highest edge alignment quality at all scales. Removing SCA results in the largest degradation, demonstrating its key role in maintaining spatial consistency and recovering fine-grained edges. HCC provides complementary improvements by enhancing hierarchical consistency, but its influence is comparatively weaker.

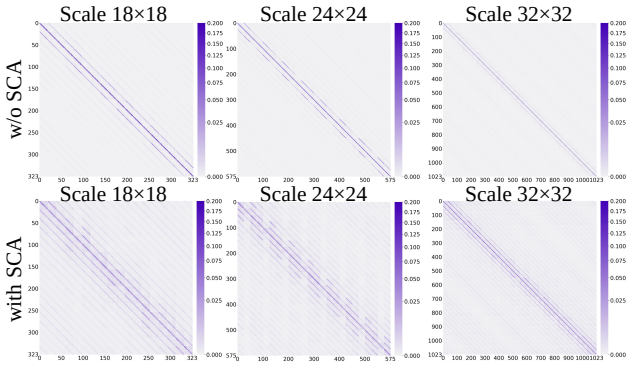

Figure 5. Impact of Spatial Consistency Autoregression (SCA) on attention maps. The attention range is significantly broader when SCA is applied.

8. Effect of Hyperparameter

We investigate the effect of the hyperparameter λ on the model's performance. Specifically, we analyze the impact of four different values of λ: 0.5, 1.0, 1.5, and 2.0, on the loss and accuracy across different scales. The results are shown in Fig. 6. As observed, when λ is small (e.g., 0.5), the model converges faster, but the token prediction accuracy is lower, particularly for the middle scales such as 6×6 and 9×9. In contrast, when λ is set to 1.0, the model achieves the highest accuracy, with a more fast convergence. As λ increases further (1.5 and 2.0), the loss converges more slowly, and the accuracy starts to decrease, suggesting that higher values of λ may hinder the model's ability to reach optimal performance for these scales. In summary, the value of λ plays a crucial role in balancing convergence speed and accuracy. A λ value of 1.0 provides the best trade-off between fast convergence and high accuracy.

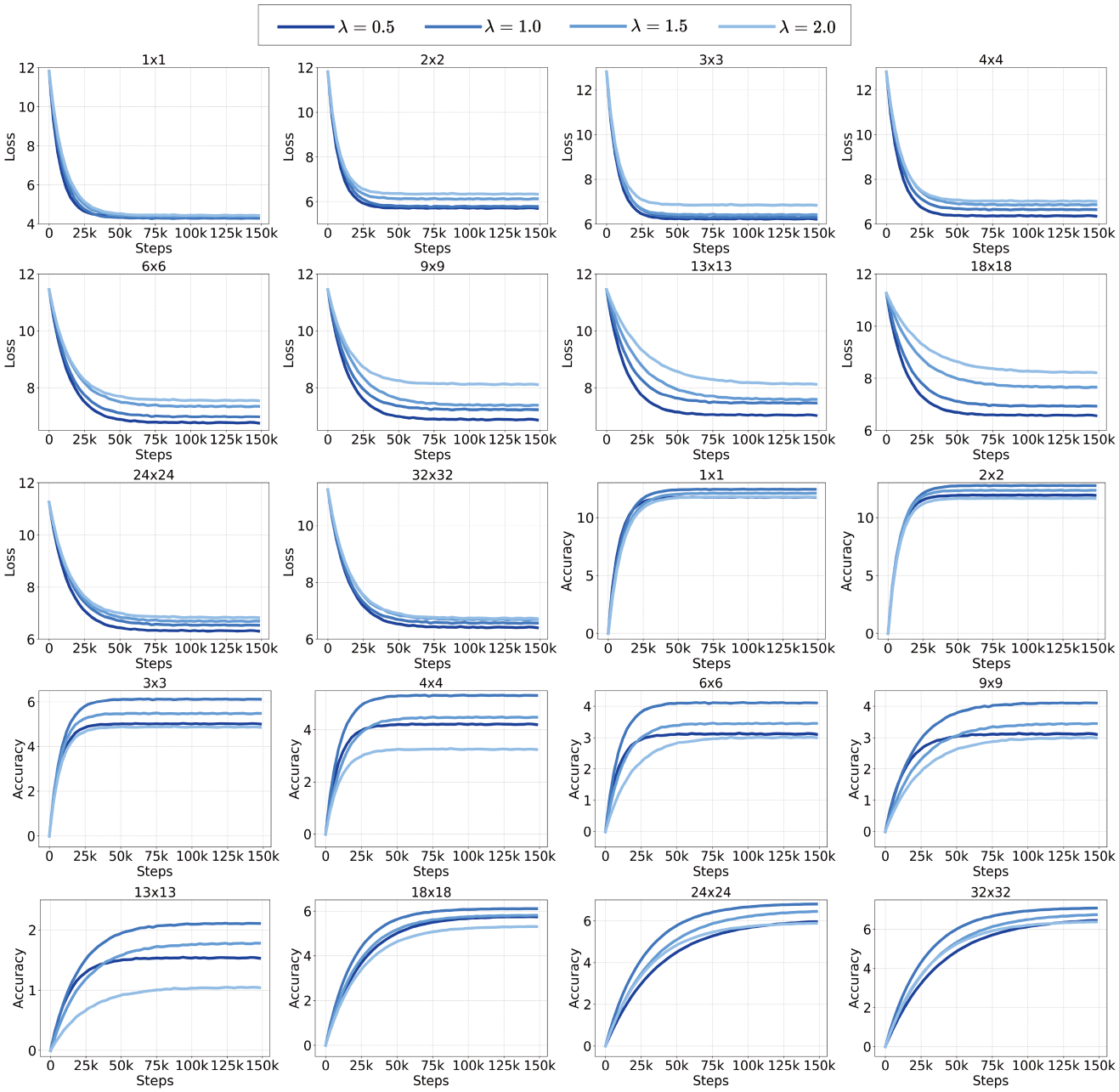

Figure 6. Effect of the hyperparameter λ on training dynamics. We compare four values of λ (0.5, 1.0, 1.5, 2.0) in terms of loss and token prediction accuracy across different scales. A smaller λ leads to faster convergence but lower accuracy, particularly at the middle scales (e.g., 6×6 and 9×9), while $\lambda = 1.0$ achieves the best overall accuracy with stable convergence. Larger λ values slow down loss convergence and degrade accuracy, indicating that excessively large consistency constraints hinder optimal learning.

9. Additional Visual Results

In this section, we present additional visual results to demonstrate the effectiveness of AlignVAR across various images. The results in Fig. 8 and Fig. 9 highlight the recovery of fine details in different scenes, with comparisons between AlignVAR and other state-of-the-art models, including GAN-based models (Real-ESRGAN [13], BSRGAN [19]) and diffusion-based models (PASD [17], UPSR [20]).

In Fig. 8, AlignVAR performs notably better in recovering intricate details such as the beak of the penguin, architectural structures. GAN-based models like BSRGAN, in contrast, tend to generate overly smooth results, losing fine

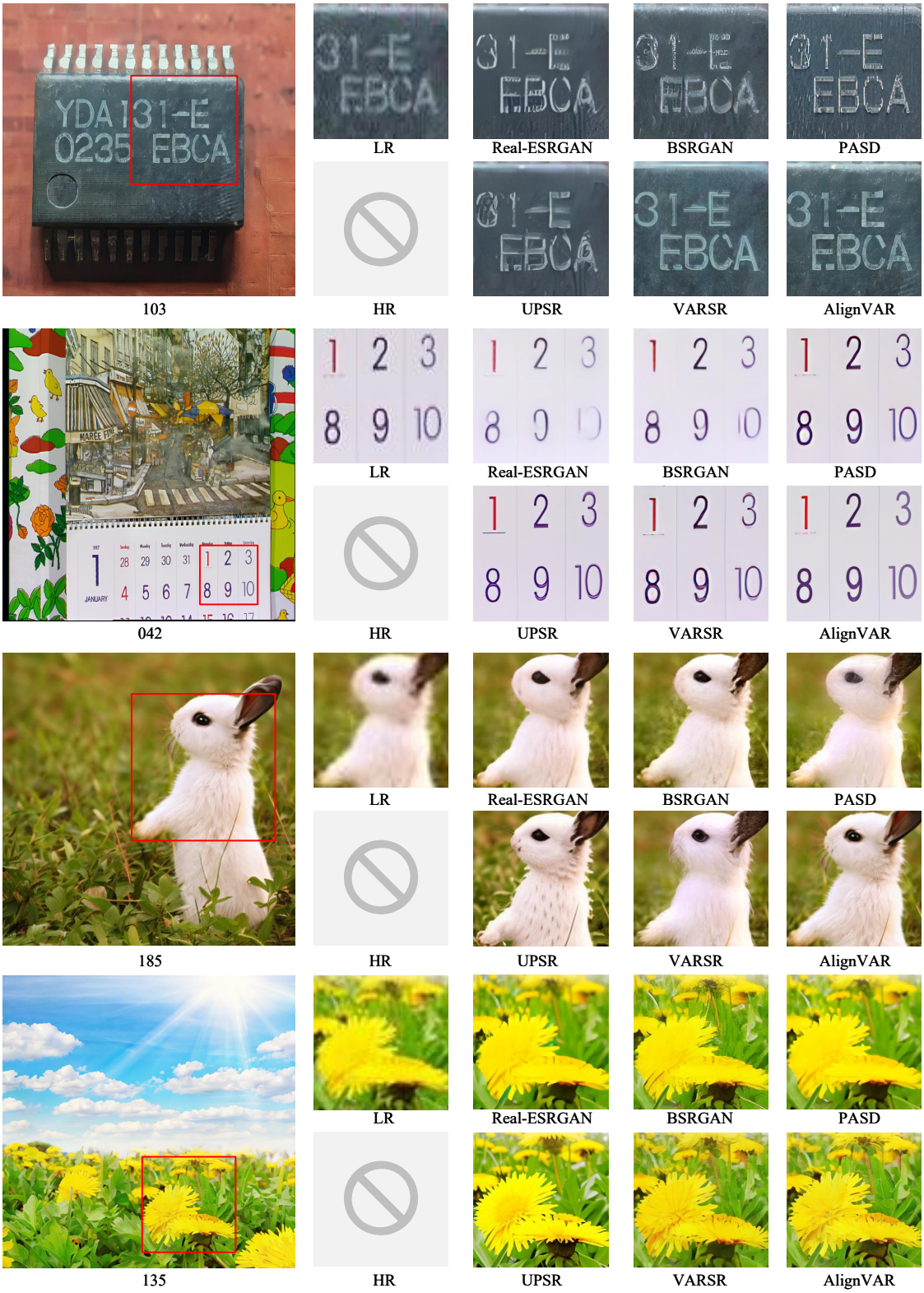

Figure 7. Comparison with state-of-the-art methods on the RealLR200 dataset. The large image on the left is the output of AlignVAR, while the small patches on the right show enlarged crops from the LR input and competing models.

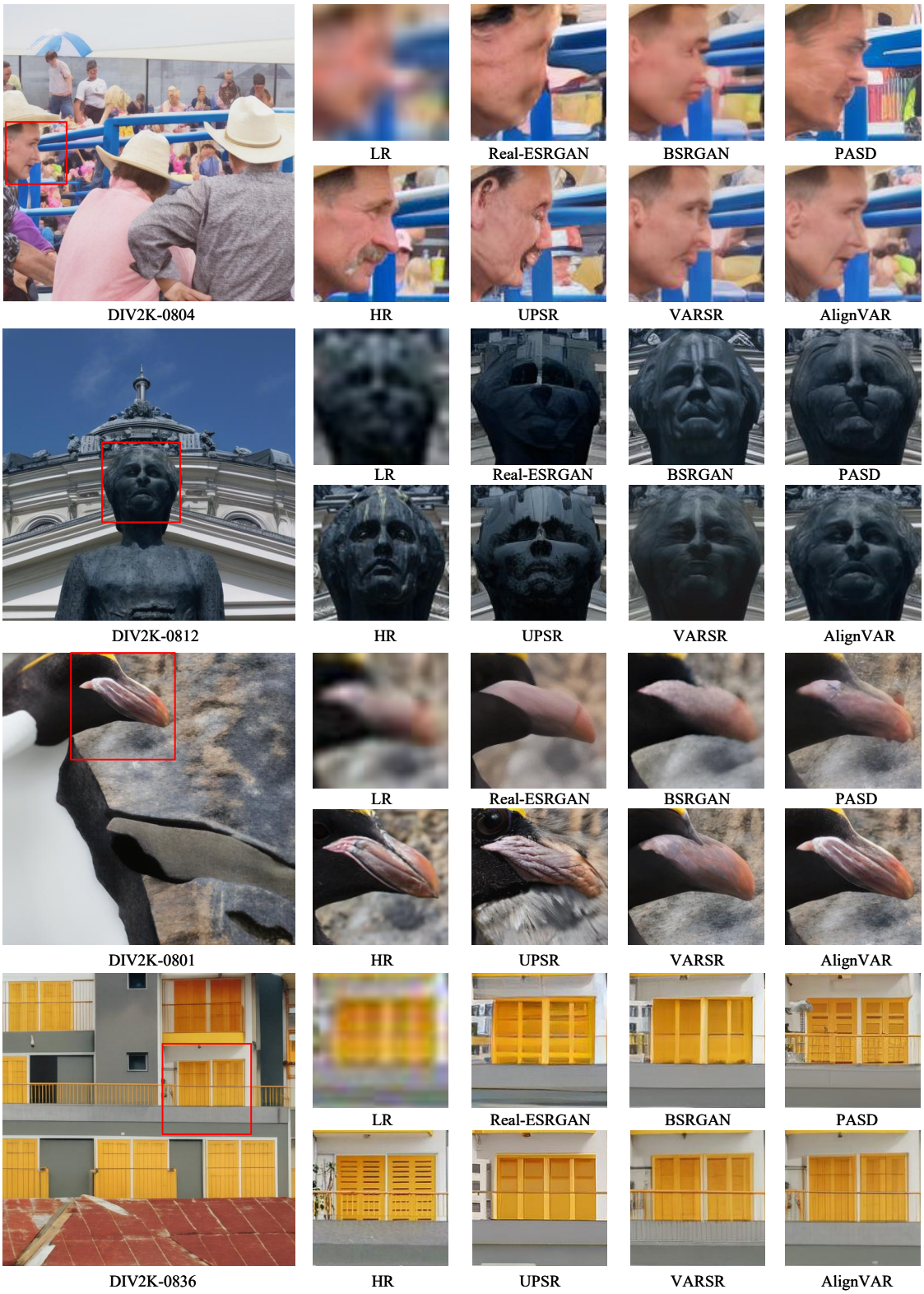

Figure 8. Additional qualitative comparisons. The large image on the left shows the output of AlignVAR, while the small patches on the right present enlarged crops from the LR input and competing methods.

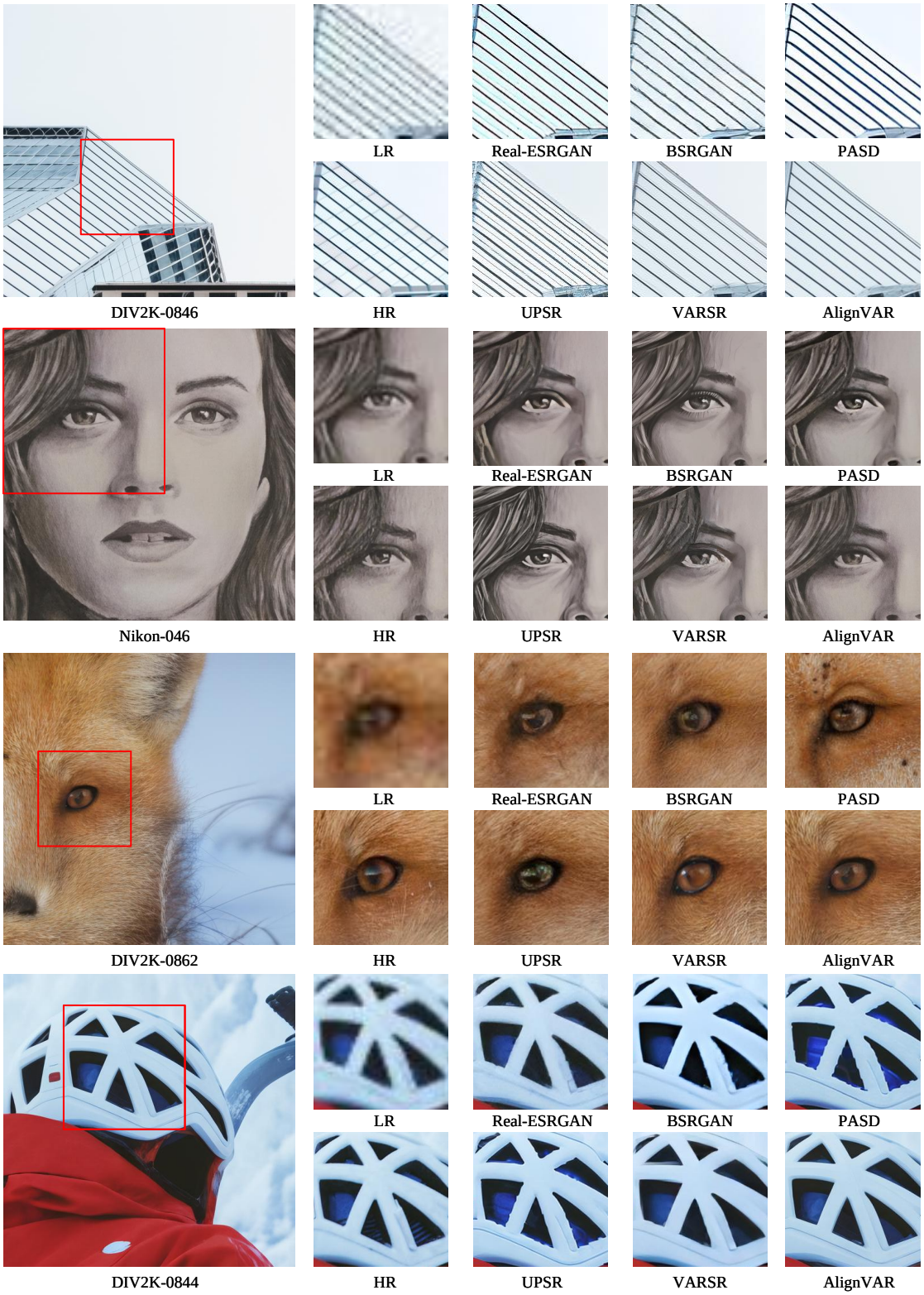

Figure 9. Additional qualitative comparisons. The large image on the left shows the output of AlignVAR, while the small patches on the right present enlarged crops from the LR input and competing methods.

details, especially in intricate structures like the penguin's beak or the architectural details. Diffusion models such as PASD also struggle with fine texture recovery, often resulting in overly blurred outputs, especially in high-frequency areas. However, compared to the HR image, some gaps remain, especially in areas with complex texture. This is likely due to the difficulty in recovering fine details, such as small textures in challenging low-resolution inputs. In Fig. 9, AlignVAR excels in recovering sketch-like images and animal details, such as the eye and fur texture of the animal. The eyes are restored with sharpness, natural shine, and detailed textures that are closer to the ground truth compared to other methods. GAN models like BSRGAN tend to oversmooth such textures, leading to unnatural results, while diffusion models again introduce unnecessary blurring, especially in small, detailed areas like the eyes or fur.

10. Limitation

While AlignVAR achieves strong performance across a wide range of benchmarks, several limitations remain. First, the recovery of middle scales proves to be particularly challenging, as shown in our hyperparameter analysis 8. These intermediate resolutions more sensitive to hyperparameter choices such as λ. In future work, we plan to explore dynamically adjustable consistency constraints that adapt to the difficulty of each scale. Second, although AlignVAR improves the spatial and hierarchical consistency of autoregressive prediction, its computation is still dominated by the final high-resolution scale, inheriting the $O(n^4)$ complexity of VAR-based models. This may limit scalability to extremely high-resolution image generation. Finally, our method relies on the quality of VQ tokenization. When the codebook fails to accurately represent rare textures or subtle structures, even consistent autoregression cannot fully recover the missing details. Incorporating stronger generative priors or adaptive codebook refinement may further enhance robustness to challenging degradations.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR workshops*, pages 126–135, 2017. 2

[2] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, pages 5128–5137, 2021. 3

[3] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *ICCV*, pages 1833–1844, 2021. 3

[4] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *ECCV*, pages 430–448. Springer, 2024. 2, 3

[5] Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Signal Processing Letters*, 19(12):813–816, 2012. 3

[6] Yunpeng Qu, Kun Yuan, Jinhua Hao, Kai Zhao, Qizhi Xie, Ming Sun, and Chao Zhou. Visual autoregressive modeling for image super-resolution. In *ICML*, 2025. 2, 3

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2, 3

[8] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *NeurIPS*, 37:84839–84865, 2024. 2

[9] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017. 1

[10] Jie Wang, Yixin Li, Yifan Wang, Siyuan Chen, Qifan Wang, and Zhangyang Wang. Clip-iqa: No-reference image quality assessment with contrastive pre-training. In *CVPR*, pages 1502–1511, 2023. 2, 3

[11] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024. 2, 3

[12] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024. 2

[13] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCV*, pages 1905–1914, 2021. 3, 5

[14] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *ECCV*, pages 101–117. Springer, 2020. 2

[15] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, pages 25456–25467, 2024. 3

[16] Sidi Yang, Tianhe Wu, Shuwei Shi, Han Gong, Jing Cao, Bin Yang, Xu Jia, and Xiaogang Zhu. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *CVPR*, pages 1191–1200, 2022. 3

[17] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *ECCV*, pages 74–91. Springer, 2024. 2, 3, 5

[18] Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *ACM MM*, page 516–520, New York, NY, USA, 2016. Association for Computing Machinery. 3, 4

[19] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, pages 4791–4800, 2021. 3, 5

[20] Leheng Zhang, Weiyi You, Kexuan Shi, and Shuhang Gu. Uncertainty-guided perturbation for image super-resolution diffusion model. In *CVPR*, pages 17980–17989, 2025. 2, 3, 5