FROM GENERATION TO CORRECTION: AUTOREGRESSIVE BITWISE CORRECTION FOR IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual autoregressive (VAR) models have achieved remarkable progress in image generation, which has boosted their application to image super-resolution (ISR). However, existing VAR-based ISR methods tend to follow a generative paradigm conditioned on the low-quality (LQ) representation directly, without explicitly disentangling reliable content from degradation-induced discrepancies. This work focuses on bitwise VAR models that operate by quantizing latent features into binary codes. Within this binary latent space, we systematically analyze the bitwise behavior between LQ and high-quality (HQ) representations. We observe that LQ and HQ representations are highly similar at the bit level, motivating us to revisit ISR through an LQ-anchored *bitwise correction* formulation. This formulation aims to identify and correct LQ bits that disagree with their HQ counterparts while preserving consistent ones. This selective correction enables faithful restoration by leveraging reliable LQ information and concentrating the generative prior on degradation-induced discrepancies for high-frequency detail synthesis. Building on this insight, we propose **Generative-Prior-enhanced Autoregressive Bitwise Correction** (**GPABC**), a framework that autoregressively predicts LQ-relative flip masks by leveraging corrected representations from preceding scales as context alongside a pretrained VAR prior. To balance reconstruction fidelity and perceptual quality, we further introduce a scale-adaptive asymmetric focal flip loss. This loss penalizes false flips at coarse scales to preserve reliable LQ content, while penalizing missed flips at fine scales to recover high-frequency details with the generative prior. Extensive quantitative and qualitative experiments across four benchmarks show that GPABC outperforms state-of-the-art methods in perceptual quality while maintaining competitive reconstruction fidelity, validating the effectiveness of the proposed bitwise correction formulation for ISR.

1 INTRODUCTION

Image super-resolution (ISR) aims to reconstruct a high-quality (HQ) image from a degraded low-quality (LQ) observation, which inherently retains reliable low-frequency content while suffering from degradation-induced discrepancies and lost high-frequency details. An effective ISR model should therefore preserve the reliable content while restoring the degraded or missing information, thereby balancing reconstruction fidelity and perceptual quality.

The development of generative models has significantly advanced ISR. Generative Adversarial Network (GAN)-based methods (Ledig et al., 2017; Wang et al., 2018; Zhang et al., 2021) improve perceptual quality but often suffer from training instability and unnatural artifacts. Diffusion Model (DM)-based methods (Wang et al., 2024; Wu et al., 2024b; Yu et al., 2024) exploit natural image priors learned from large-scale data and achieve impressive visual quality. However, iterative sampling incurs high computational costs, while accelerated one-step variants (Wu et al., 2024a; Kim et al., 2026; Fang et al., 2026) still struggle to balance realistic detail synthesis with faithful reconstruction.

Visual autoregressive (VAR) modeling (Tian et al., 2024) provides a promising alternative for ISR. Through next-scale prediction, VAR models generate images from coarse structure to fine detail. Infinity (Han et al., 2025) extends VAR by using binary spherical quantization (BSQ) to quantize

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

latent features into binary bits. Recent methods (Qu et al., 2025; Wei et al., 2025; Kong et al., 2025; Zhu et al., 2026; Zheng et al., 2026) have applied the VAR paradigm to ISR. However, existing methods follow a generative paradigm that directly conditions on the LQ representation through global signals (Liu et al., 2026; Hadji et al., 2026) or prefix tokens (Kong et al., 2025; Zheng et al., 2026), without explicitly disentangling reliable LQ content from degradation-induced errors. We therefore revisit ISR by retaining reliable content while correcting discrepancies.

In this work, we focus on bitwise VAR models, exemplified by Infinity, which perform image generation via multi-scale bitwise prediction. We systematically analyze the bitwise behaviors of LQ and HQ representations in their discrete latent space. By comparing their binary representations, we make four key observations: (1) LQ and HQ representations are highly similar at the bit level, particularly at coarse scales; (2) degradation-induced bit changes exhibit structured and learnable patterns rather than random variations; (3) quantization confidence strongly correlates with bit agreement between LQ and HQ representations; and (4) the degradation sensitivity of bit channels exhibits distinct patterns across scales. These observations motivate us to revisit ISR through an LQ-anchored ***bitwise correction*** formulation that learns to identify and correct LQ bits differing from their HQ counterparts while preserving those already in agreement.

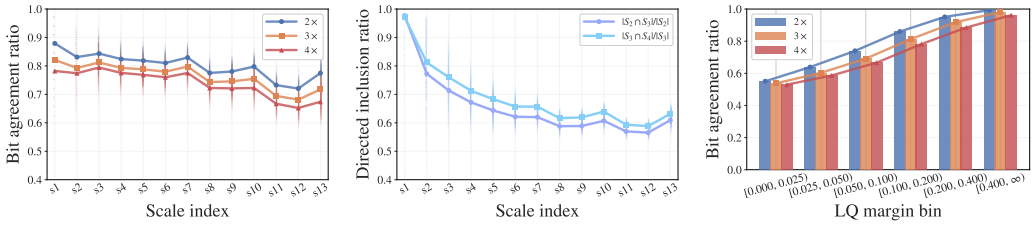
To this end, we propose **Generative-Prior-enhanced Autoregressive Bitwise Correction** (**GPABC**), a framework that leverages generative priors to predict bitwise corrections for LQ representations and progressively align them with their HQ counterparts. Specifically, at each scale, GPABC takes the corrected representations from preceding scales as context and employs the generative prior of a pretrained bitwise VAR model to predict an LQ-relative bit-flip mask. The predicted mask determines which LQ bits should be flipped to align with the HQ representation. In this way, GPABC explicitly preserves reliable bits that are in agreement between LQ and HQ representations while selectively correcting their disagreement bits. To balance reconstruction fidelity and perceptual quality, we further introduce a scale-adaptive asymmetric focal flip loss. At coarse scales, it assigns larger penalties to false flips, preserving reliable structural information in LQ observations from being corrupted. At fine scales, it assigns larger penalties to missed flips, encouraging the generative prior to synthesize high-frequency details. In addition, we adopt a curriculum learning strategy to mitigate exposure bias between teacher-forcing training and autoregressive inference.

Extensive experiments demonstrate that GPABC achieves state-of-the-art performance in balancing reconstruction fidelity and perceptual quality. Compared with state-of-the-art diffusion-based and VAR-based ISR methods, GPABC yields the highest MANIQA and CLIPIQA scores and top-2 TOPIQ scores across all benchmarks, while maintaining competitive full-reference performance.

2 Related Work

Image Super-Resolution. ISR addresses the ill-posed problem of recovering a high-quality image by modeling and inverting complex real-world degradations. GAN-based methods (Ledig et al., 2017; Wang et al., 2018; 2021; Zhang et al., 2021) employ adversarial learning to recover perceptually realistic details, but often suffer from training instability and limited generative priors. To address these limitations, DM-based methods (Wang et al., 2024; Wu et al., 2024b; Yu et al., 2024) leverage rich generative priors inherent in pretrained diffusion models and thus excel in perceptual quality. Recent one-step methods (Wu et al., 2024a; Sun et al., 2025; Kim et al., 2026; Fang et al., 2026) improve inference efficiency through LQ-guided initialization and streamlined sampling. Despite these advances, simultaneously achieving high perceptual quality, reconstruction fidelity, and inference efficiency remains an open challenge for diffusion-based ISR.

Visual Autoregressive Modeling. Visual autoregressive modeling (Tian et al., 2024) reformulates image autoregression as next-scale prediction, enabling high-quality generation with efficient sampling. Infinity (Han et al., 2025) further enhances generative capacity by introducing bitwise token prediction with an infinite vocabulary. VARSR (Qu et al., 2025) extends this paradigm to ISR by conditioning VAR generation on LQ-derived prefix tokens, demonstrating the potential of VAR models for efficient restoration. Subsequent methods (Wei et al., 2025; Kong et al., 2025; Zhu et al., 2026) further explore this paradigm. DVAR (Zheng et al., 2026) introduces canonical scale scheduling to support size-flexible restoration. AlignVAR (Liu et al., 2026) tailors VAR to ISR by exploiting spatial and hierarchical consistency, while HIT (Hadji et al., 2026) reduces model complexity by integrating hierarchical image tokenization with preference optimization over LQ–HQ pairs. Despite



(a) Scale-wise bit agreement. (b) Directed flip inclusion. (c) Margin-bit relationship.

Figure 1: Analysis of binary bit-level behavior between LQ and HQ representations. (a) Scale-wise bit agreement ratio under different degradation levels. (b) Directed inclusion ratios between flip sets across degradation levels. (c) The relationship between bit margin and agreement ratio.

these advances, existing VAR-based ISR methods largely retain a generation-oriented formulation, treating LQ information as conditioning signals or prefix tokens rather than modeling the shared information between LQ and HQ representations.

3 Analysis

Based on the bitwise VAR model Infinity (Han et al., 2025), we systematically analyze the bit-level behavior of LQ images and their HQ counterparts in its discrete representation space. We sample 2,000 image groups from the LSDIR (Li et al., 2023) training dataset, each comprising one high-resolution HQ image and its corresponding $2\times$, $3\times$, and $4\times$ low-resolution degraded images. All images are resized and center-cropped to 1024×1024, then encoded by the Infinity BSQ tokenizer into binary bits at 13 scales.

Scale-wise bit agreement under different degradation levels. To investigate the bit-level relationship between LQ and HQ images, we measure their bit agreement ratio at each scale under different degradation levels. As shown in Fig. 1a, their binary bit representations remain highly similar. At scale 1, the agreement ratios are 87.9%, 82.1%, and 78.2% under $2\times$, $3\times$, and $4\times$ degradations, respectively. As the scale increases, the bit agreement ratio decreases with fluctuations. Moreover, stronger degradations consistently yield lower bit agreement at each scale. Nevertheless, the lowest bit agreement ratio remains above 65%. Thus, degradation-induced disagreements occur only on a small fraction of bits, leaving most LQ and HQ bits in agreement. Since discrete bits are binary, aligning LQ with HQ reduces to locating and flipping the few disagreeing bits while preserving the rest. This motivates us to formulate ISR as bitwise correction via LQ-anchored flip-mask prediction, keeping the already agreed bits unchanged and correcting the disagreed bits to match the HQ targets.

Directed inclusion of flip sets across degradation levels. We further quantify the directed inclusion between the sets of flipped bits under different degradation levels. Let S_r denote the set of LQ bits that disagree with their HQ counterparts under $r\times$ degradation. We evaluate the directed inclusion ratio $|S_a \cap S_b|/|S_a|$ for $a < b$. As shown in Fig. 1b, at scale 1, 98% of the flips under $2\times$ degradation persist under $3\times$ degradation, while 97% of those under $3\times$ degradation persist under $4\times$ degradation. Although the inclusion ratio gradually declines with minor fluctuations toward finer scales, it remains above 55% across all 13 scales. Under the assumption of independent random flipping, the theoretical expected value is merely 17.9% for $|S_2 \cap S_3|/|S_2|$ and 21.8% for $|S_3 \cap S_4|/|S_3|$. These results demonstrate that stronger degradation largely inherits the bits affected by milder degradation, confirming that bit flips do not occur independently and randomly, but instead exhibit degradation-dependent, learnable, and structured characteristics.

Relationship between bit margin and bit agreement. To investigate the relationship between quantization confidence and bit flips, we compute the distance from each entry of the normalized LQ feature to the sign-quantization boundary as its bit margin, bin the bits accordingly, and measure the bit agreement within each bin. Fig. 1c shows that the bit agreement ratio exhibits a steady, monotonic increase with larger margins across all degradation levels. For bits with margins below 0.025, the agreement ratios are merely 55.0%, 53.8%, and 52.9% under $2\times$, $3\times$, and $4\times$ degradations,

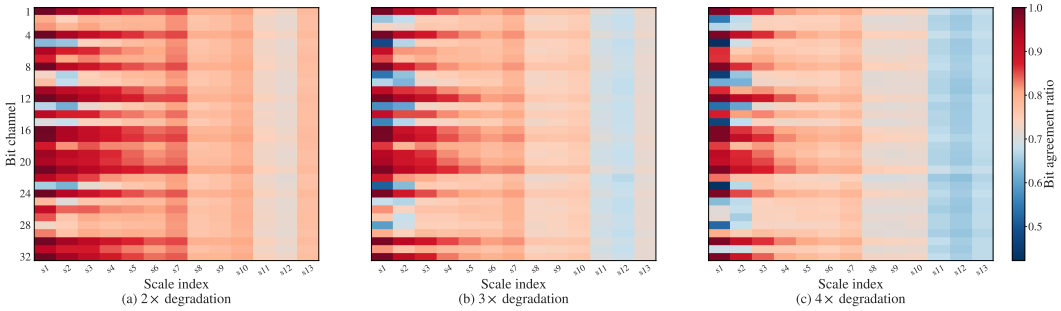


Figure 2: Channel-wise degradation sensitivity at each scale under different degradations.

respectively. In contrast, for bits with margins of at least 0.4, the agreement ratios rise significantly to 99.4%, 98.1%, and 96.2%. Low-margin bits lie near the quantization boundary where degradation perturbations readily alter their signs, leading to high uncertainty, whereas high-margin bits exhibit remarkable stability against degradation. Therefore, the margin can serve as an effective prior for evaluating LQ bit reliability, guiding the model to preserve high-margin bits while focusing its correction capacity on uncertain low-margin ones.

Channel-wise degradation sensitivity. We further compute channel-wise bit agreement between LQ representations and their HQ counterparts to examine channel sensitivity to degradation. As shown in Fig. 2, coarse scales contain both highly stable and sensitive channels, whereas almost all channels become sensitive at finer scales. Specifically, at scale 1, channels $\{1, 4, 8, 12, 16, 17, 21, 24, 30, 32\}$ maintain agreement above 98% under all degradation levels, while channels 5 and 23 flip frequently. This reveals a strong structural correlation between coarse-scale bits and degradation. We believe that individual channels at coarse scales encode distinct low-frequency visual attributes, with some primarily encoding degradation-invariant properties such as color. After approximately scale 7, this channel-specific pattern weakens. At finer scales, bit flips become more unstable, as channels describe localized high-frequency details, which are more sensitive to degradation perturbations. This suggests a scale-adaptive correction strategy that heavily references LQ inputs at coarse scales to exploit stable structures, and progressively shifts toward generative-prior-enhanced correction at finer scales to recover high-frequency details lost to degradation.

4 METHOD

4.1 PRELIMINARY: BITWISE VISUAL AUTOREGRESSIVE MODEL

Next-Scale Prediction. Visual autoregressive models adopt a scale-wise prediction paradigm for image generation. Given an image I, a visual encoder first maps it into a continuous feature representation $F \in \mathbb{R}^{h\times w\times d}$, which is then decomposed into K residual token maps $(R_1, R_2, \ldots, R_K)$. Their joint distribution is factorized as

$$p(R_1, \ldots, R_K) = \prod_{k=1}^{K} p(R_k \mid R_1, \ldots, R_{k-1}, c), \tag{1}$$

where c denotes the text embeddings injected through cross-attention. This next-scale prediction paradigm generates images hierarchically from coarse structures to fine details.

Binary Spherical Quantization. In this work, we focus on Infinity, which replaces conventional vector quantization with binary spherical quantization (BSQ), scaling the vocabulary size to 2^d, where d denotes the code embedding dimension. At scale k, the continuous residual feature map $Z_k \in \mathbb{R}^{h_k\times w_k\times d}$ is computed as

$$Z_k = \mathrm{down}(F - F_{k-1},\ (h_k, w_k)), \tag{2}$$

where $\mathrm{down}(\cdot)$ denotes downsampling and (h_k, w_k) is the spatial size at scale k. The cumulative feature $F_{k-1} \in \mathbb{R}^{h\times w\times d}$ is obtained by upsampling and summing the quantized residual maps from

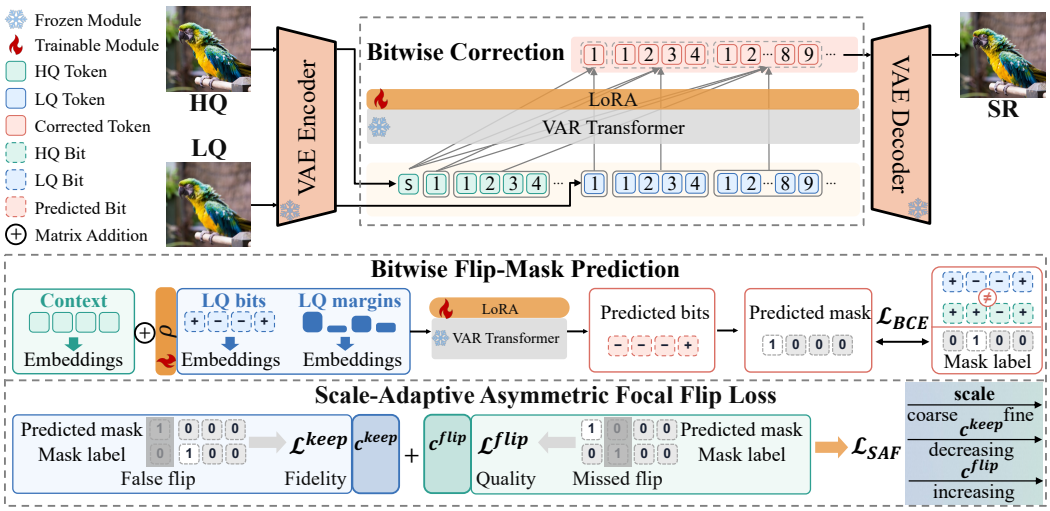


Figure 3: Overview of GPABC. Given LQ and HQ images, the frozen VAE encoder maps them into continuous features, which are then quantized into multi-scale tokens. At each scale, corrected bits from preceding scales form the autoregressive context, while the current LQ bits and margins, weighted by a learnable parameter ρ, are fed into the Transformer equipped with trainable LoRA adapters. The predicted bit logits yield an LQ-relative flip mask, supervised by the disagreement between LQ and HQ bits through $\mathcal{L}_{\mathrm{BCE}}$. The scale-adaptive asymmetric focal flip loss $\mathcal{L}_{\mathrm{SAF}}$ penalizes false flips with a scale-decreasing weight c^{keep} and missed flips with a scale-increasing weight c^{flip}. The corrected multi-scale tokens are finally decoded into the super-resolution (SR) image.

the preceding scales, with $F_0 = \mathbf{0}$. Let $z_{k,i} \in \mathbb{R}^d$ denote the feature at the i-th spatial position of Z_k, where $i \in \{1, \ldots, h_k w_k\}$. Its binary code is obtained by normalization and BSQ quantization:

$$q_k(i,j) = [\mathcal{Q}(z_{k,i})]_j = \frac{1}{\sqrt{d}} \operatorname{sign}\left(\frac{[z_{k,i}]_j}{|z_{k,i}|}\right), \tag{3}$$

where $q_k(i,j) \in \{-\frac{1}{\sqrt{d}}, +\frac{1}{\sqrt{d}}\}$ denotes the value of the j-th bit channel at position i of scale k.

4.2 GPABC Framework

Motivated by the analysis in Sec. 3, we propose **Generative-Prior-enhanced Autoregressive Bitwise Correction** (**GPABC**), a framework that leverages generative priors to correct LQ binary bits from coarse to fine. The pipeline of our GPABC is illustrated in Fig. 3. At each scale, GPABC uses previously corrected bits as autoregressive context and the current LQ bits as the correction anchor to predict a bit-flip mask, whose label is derived from the bitwise disagreement between the LQ and HQ. In addition, a scale-adaptive asymmetric focal flip loss balances coarse-scale fidelity and fine-scale perceptual quality. Finally, the corrected multi-scale representation is decoded into the SR image.

Shared-Prefix Bit and Margin Computation. To focus the model on learning the bitwise correction required at the current scale, we compute the LQ and HQ bits under a shared cumulative prefix. Given the continuous features F^{LQ} and F^{HQ} extracted by the frozen VAE encoder, the cumulative prefix feature is reconstructed from the quantized residual maps of the preceding $k-1$ scales:

$$F_{k-1} = \sum_{l=1}^{k-1} \mathrm{up}\Big(\widetilde{R}_l,\ (h,w)\Big), \qquad F_0 = \mathbf{0}, \tag{4}$$

where $\mathrm{up}(\cdot)$ denotes the upsampling operation. During teacher-forcing, $\widetilde{R}_l$ is the quantized residual map of the HQ image. During curriculum rollout, $\widetilde{R}_l$ is constructed from HQ bits and detached predicted corrected bits. Using this shared prefix, the continuous LQ and HQ residual maps at scale k are computed as

$$Z_k^x = \mathrm{down}(F^x - F_{k-1},\ (h_k, w_k)), \qquad x \in \{\mathrm{LQ}, \mathrm{HQ}\}. \tag{5}$$

After extracting the continuous residual feature maps Z_k^{LQ} and Z_k^{HQ} at scale k, we follow the BSQ procedure in Sec. 4.1 to obtain their corresponding discrete binary representations q_k^{LQ} and q_k^{HQ}, respectively. Furthermore, we define m_k^{LQ} as the absolute values of the channel-wise normalized continuous features, with each component measuring its distance to the BSQ sign boundary.

Bitwise Flip-Mask Prediction. We learn to correct LQ bits using the frozen Infinity Transformer equipped with trainable LoRA adapters, thereby exploiting its pretrained generative prior to enhance the correction process. Since the BSQ outputs are binary, bitwise correction can be implemented by predicting whether each LQ bit should be preserved or flipped. As shown in Sec. 3, a larger margin indicates that the corresponding LQ bit lies farther from the quantization boundary and is therefore more reliable. Therefore, the model takes the current-scale LQ bits and their quantization margins, together with the corrected prefix from the preceding $k-1$ scales, as inputs. Since coarse-scale LQ bits are generally more reliable for preserving global structure, whereas fine-scale bits encode more high-frequency details, we introduce a learnable scale-wise weight ρ_k to adaptively control the LQ conditioning strength:

$$e_k = \rho_k \left[f_{\mathrm{bit}}\left(q_k^{\mathrm{LQ}}\right) + f_{\mathrm{margin}}\left(m_k^{\mathrm{LQ}}\right) \right], \tag{6}$$

where ρ_k is initialized to 1 for each scale. f_{bit} and f_{margin} are learnable projections into the Transformer hidden space. The LQ conditioning embedding e_k is added to the embedding derived from the preceding $k-1$ corrected scales and fed into the Transformer equipped with LoRA adapters. Based on the restored low-frequency content encoded by the corrected prefix, as well as the bit reliability indicated by the margins, the model, enhanced by the pretrained generative prior, determines whether each LQ bit should be preserved or flipped. Let l_k^0 and l_k^1 denote the logits for negative and positive bit values, respectively. The flip probability can be obtained by reparameterizing the HQ-bit prediction as an LQ-relative prediction:

$$\pi_k = \sigma\left[\mathrm{sign}\left(q_k^{\mathrm{LQ}}\right) \odot \left(l_k^0 - l_k^1\right) \right], \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $\odot$ denotes element-wise multiplication. With the bit-flip label $M_k = \mathbb{I}\left[q_k^{\mathrm{LQ}} \neq q_k^{\mathrm{HQ}}\right]$, where $\mathbb{I}[\cdot]$ denotes the indicator function, the optimization objective is

$$\mathcal{L}_{\mathrm{BCE}} = -\sum_{k=1}^{K} \sum_{(i,j)} \frac{\omega_k}{dN_k} \Big[M_k(i,j) \log \pi_k(i,j) + (1 - M_k(i,j)) \log(1 - \pi_k(i,j)) \Big], \tag{8}$$

where $N_k = h_k w_k$ and $\omega_k = \frac{\sqrt{N_k}}{\sum_{l=1}^{K} \sqrt{N_l}}$ denotes the scale weight, which balances the contributions of coarse and fine scales with different numbers of tokens.

Scale-Adaptive Asymmetric Focal Flip Loss. We further design a scale-adaptive asymmetric focal flip loss to adjust the optimization emphasis across scales. As shown by the bit agreement analysis in Sec. 3, LQ and HQ bits exhibit higher agreement at coarse scales. The model should therefore penalize false flips more strongly to preserve reliable LQ bits at coarse scale. At finer scales, the agreement decreases as the bits increasingly represent degraded high-frequency details. The model should thus place greater emphasis on missed flips to encourage detail generation. We apply focal modulation to the bitwise predictions and define $\mathcal{L}_k^{\mathrm{keep}}$ to penalize bits that should be preserved but are incorrectly predicted to be flipped:

$$\mathcal{L}_k^{\mathrm{keep}} = -\frac{\sum_{(i,j)} (1 - M_k(i,j)) \, (\pi_k(i,j))^{\gamma} \log (1 - \pi_k(i,j))}{\sum_{(i,j)} (1 - M_k(i,j))}, \tag{9}$$

where γ is the focal modulation factor. To penalize bits that should be flipped but are incorrectly predicted to be preserved, we define

$$\mathcal{L}_k^{\mathrm{flip}} = -\frac{\sum_{(i,j)} M_k(i,j) \, (1 - \pi_k(i,j))^{\gamma} \log (\pi_k(i,j))}{\sum_{(i,j)} M_k(i,j)}. \tag{10}$$

To adaptively balance these two correction behaviors across scales, we define scale-dependent weights for $\mathcal{L}_k^{\text{keep}}$ and $\mathcal{L}_k^{\text{flip}}$ as

$$c_k^{\text{keep}} = 1 + \alpha(1 - r_k)^2, \qquad c_k^{\text{flip}} = 1 + \alpha r_k^2, \tag{11}$$

where α controls the degree of asymmetry and $r_k = (k-1)/(K-1)$ denotes the normalized scale index. As the scale index increases, c_k^{keep} decreases, whereas c_k^{flip} increases. The resulting scale-adaptive asymmetric focal flip loss is

$$\mathcal{L}_{\text{SAF}} = \sum_{k=1}^{K} \omega_k \frac{c_k^{\text{keep}} \mathcal{L}_k^{\text{keep}} + c_k^{\text{flip}} \mathcal{L}_k^{\text{flip}}}{c_k^{\text{keep}} + c_k^{\text{flip}}}. \tag{12}$$

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \lambda \mathcal{L}_{\text{SAF}}, \tag{13}$$

where λ controls the contribution of $\mathcal{L}_{\text{SAF}}$ during training.

Curriculum Learning. We adopt a two-stage training scheme organized as a progressive prefix curriculum. In the first stage, we employ teacher-forcing training, where HQ bits of the preceding scales form an error-free prefix, allowing the model to learn bit flip prediction under a clean multi-scale context. Meanwhile, $\mathcal{L}_{\text{SAF}}$ is disabled so that the model can first establish a reliable basic correction capability using $\mathcal{L}_{\text{BCE}}$. In the second stage, $\mathcal{L}_{\text{SAF}}$ is activated, and detached predicted corrected bits progressively replace HQ bits at disagreement positions with an increasing probability. This rollout gradually exposes the model to its own predictions without propagating gradients through the discrete corrections.

5 Experiments

5.1 Experimental Setups

Datasets. The training data consists of 30K high quality images collected from web sources through a careful quality filtering process. The corresponding text annotations are generated by LLaVA (Liu et al., 2024). We use the standard degradation pipeline of Real-ESRGAN (Wang et al., 2021) to synthesize the LQ-HQ training pairs. Following previous work (Kong et al., 2025), the evaluation data comprises a synthetic benchmark DIV2K (Agustsson & Timofte, 2017) and three real-world benchmarks RealSR (Cai et al., 2019), DRealSR (Wei et al., 2020) and RP60 (Yu et al., 2024).

Training Details. GPABC is trained for $\times 4$ ISR, targeting a final resolution of 1024×1024. We utilize LoRA to inject trainable parameters into the cross-attention, self-attention, and feed-forward layers of an Infinity-2B backbone, with the rank set to 8. The model is trained with a batch size of 16, a learning rate of 7.5×10^{-4} and a weight decay of 10^{-2} using AdamW optimizer. GPABC is trained for 22K iterations. The first 2K iterations use teacher-forcing with λ in Eq. (13) set to 0. During the remaining 20K iterations, predicted and HQ bits are mixed at their disagreement positions, with the probability of using predictions increasing linearly to 1. Meanwhile, λ increases linearly from 0 to 1 over the first 1K iterations of this stage and remains fixed thereafter. The hyperparameters γ and α are both set to 2.0. All experiments are conducted on 16 NVIDIA H20 GPUs.

Metrics. Considering the trade-off between fidelity and perceptual quality, we adopt both full-reference and no-reference Image Quality Assessment (NR-IQA) metrics for comprehensive evaluation. The full-reference metrics include PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), and DISTS (Ding et al., 2020), while the NR-IQA metrics include MANIQA (Yang et al., 2022), CLIPIQA (Wang et al., 2023), MUSIQ (Ke et al., 2021), and TOPIQ (Chen et al., 2024).

5.2 Comparison with SOTA

We compare GPABC with state-of-the-art (SOTA) methods in two categories to conduct both quantitative and qualitative comparisons. The DM-based models include StableSR (Wang et al., 2024), SeeSR (Wu et al., 2024b), SUPIR (Yu et al., 2024), OSEDiff (Wu et al., 2024a), PiSA-SR (Sun et al., 2025), while the AR-based methods include VARSR (Qu et al., 2025), PURE (Wei et al., 2025), NSARM (Kong et al., 2025), DVAR (Zheng et al., 2026) and VARestorer (Zhu et al., 2026).

Table 1: Comparison with SOTA methods on synthetic and real-world benchmarks. The best and second-best performances are highlighted in red and blue, respectively.

Dataset	Metrics	DM-based					AR-based					
		StableSR	SeeSR	SUPIR	OSEDiff	PiSA-SR	VARSR	VARestorer	PURE	DVAR	NSARM	GPABC
DIV2K-Val	PSNR↑	20.13	20.63	19.92	20.40	20.37	20.70	20.26	18.82	19.67	20.03	20.53
	SSIM↑	0.5052	0.5135	0.4792	0.5170	0.5096	0.5094	0.4868	0.4489	0.4805	0.4754	0.5095
	LPIPS↓	0.2973	0.2854	0.2845	0.2971	0.2649	0.2758	0.2587	0.2772	0.3045	0.2668	0.2577
	DISTS↓	0.1856	0.1687	0.1615	0.1791	0.1709	0.1698	0.1844	0.1591	0.1803	0.1554	0.1646
	MANIQA↑	0.4917	0.5513	0.5051	0.4490	0.4865	0.5490	0.4330	0.5765	0.5860	0.7258	0.7486
	CLIPIQA↑	0.7365	0.7273	0.6738	0.7132	0.7274	0.7491	0.6354	0.7560	0.7534	0.8322	0.8654
	MUSIQ↑	72.17	72.78	72.36	71.97	73.08	74.40	71.89	72.80	74.24	74.06	74.18
	TOPIQ↑	0.6960	0.7324	0.6974	0.6343	0.6896	0.7324	0.6661	0.7206	0.7628	0.7721	0.7838
RealSR	PSNR↑	24.80	25.15	22.97	25.15	25.50	25.59	24.91	23.38	23.57	23.54	23.54
	SSIM↑	0.7186	0.7210	0.5515	0.7341	0.7418	0.7266	0.6906	0.6502	0.6751	0.6412	0.6396
	LPIPS↓	0.2973	0.3008	0.4816	0.2920	0.2672	0.3225	0.2733	0.3496	0.3220	0.3694	0.4069
	DISTS↓	0.2159	0.2224	0.2717	0.2128	0.2044	0.2358	0.2188	0.2348	0.2315	0.2547	0.2688
	MANIQA↑	0.4422	0.5437	0.4335	0.4717	0.4810	0.5482	0.4437	0.4495	0.5556	0.6046	0.6695
	CLIPIQA↑	0.6199	0.6703	0.5742	0.6688	0.6704	0.6990	0.6148	0.6623	0.6823	0.7356	0.7831
	MUSIQ↑	66.09	69.82	61.00	69.09	70.15	71.23	68.54	65.42	70.30	71.13	70.66
	TOPIQ↑	0.5899	0.6888	0.5651	0.6248	0.6369	0.6990	0.6175	0.5926	0.6982	0.7315	0.7266
DRealSR	PSNR↑	28.40	28.07	24.35	27.92	28.32	28.17	27.26	26.13	26.78	26.66	26.90
	SSIM↑	0.7718	0.7683	0.5222	0.7835	0.7804	0.7656	0.7276	0.6867	0.7319	0.6910	0.7014
	LPIPS↓	0.3123	0.3175	0.5676	0.2967	0.2959	0.3538	0.3184	0.3834	0.3506	0.4042	0.4301
	DISTS↓	0.2222	0.2315	0.2930	0.2162	0.2169	0.2521	0.2375	0.2493	0.2466	0.2658	0.2862
	MANIQA↑	0.3987	0.5129	0.4446	0.4657	0.4856	0.5358	0.4353	0.4205	0.5472	0.5958	0.6361
	CLIPIQA↑	0.6125	0.6912	0.6093	0.6961	0.6969	0.7249	0.6420	0.6390	0.7181	0.7226	0.7653
	MUSIQ↑	58.74	65.09	59.85	64.68	66.11	68.02	65.37	59.96	67.70	67.50	66.68
	TOPIQ↑	0.5408	0.6574	0.5815	0.6000	0.6330	0.6796	0.5987	0.5584	0.6833	0.6988	0.7036
RP60	MANIQA↑	0.5001	0.5536	0.4774	0.4813	0.5273	0.5555	0.4782	0.5689	0.6214	0.7218	0.7535
	CLIPIQA↑	0.7535	0.7588	0.7153	0.7432	0.7729	0.7679	0.7064	0.7964	0.8093	0.8554	0.8910
	MUSIQ↑	71.91	72.31	71.02	71.15	73.41	73.71	72.41	72.47	74.01	73.88	74.10
	TOPIQ↑	0.6879	0.7250	0.6636	0.6522	0.7056	0.7225	0.6904	0.7132	0.7709	0.7612	0.7716

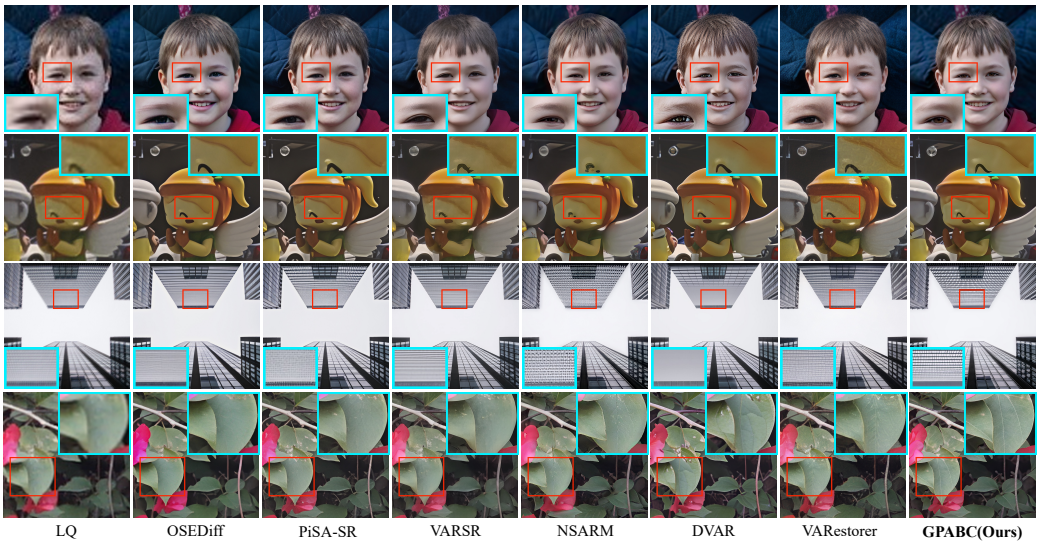


Figure 4: Qualitative Comparisons of the proposed GPABC and other SOTA methods. Please zoom in for a better view.

Quantitative Comparisons. As shown in Tab. 1, the proposed GPABC is compared with advanced DM-based and AR-based methods on four benchmark datasets. Regarding full-reference metrics, DM-based methods (*e.g.*, OSEDiff and PiSA-SR) and the AR–DM hybrid VARSR generally achieve higher PSNR/SSIM. However, their lower no-reference metric scores indicate less favorable perceptual quality. Meanwhile, several AR-based methods (*e.g.*, DVAR and NSARM) achieve higher no-reference metrics at the cost of sacrificing a certain degree of fidelity. In comparison, GPABC obtains the highest MANIQA and CLIPIQA scores, as well as the top-2 TOPIQ score, while maintaining comparable full-reference metric scores. Overall, by preserving reliable bits from LQ representations and leveraging the generative prior to synthesize high-frequency details, GPABC achieves a favorable balance between perceptual quality and reconstruction fidelity.



Figure 5: Visualization for ablation study. Please zoom in for a better view.

Qualitative Comparisons. Fig. 4 displays qualitative comparisons on four examples of four benchmarks. From the examples in the third and fourth rows of Fig. 4, it can be observed that OSEDiff, PiSA-SR and VARSR tend to generate relatively smooth restorations, where some fine-grained details, such as leaf veins and building windows, are not fully recovered. Methods such as NSARM, DVAR and VARestorer may occasionally introduce artifacts or inconsistent structures in severely degraded regions, such as the eyes of the person and the cartoon sculpture in the first and second rows. In comparison, GPABC restores sharper and more natural details while better preserving the structural consistency of the input, especially in challenging regions such as fine textures, edges, and small patterns. These visual results are consistent with the metrics reported in Tab. 1, where GPABC achieves superior perceptual scores while maintaining comparable full-reference performance.

5.3 Ablation Study

We conduct ablation experiments on the DIV2K-Val benchmark to evaluate the effectiveness of each component in GPABC. The results are shown in Tab. 2 and Fig. 5.

Impact of Scale-Wise Weight ρ. The reliability and contribution of LQ information vary across scales, motivating learnable scale-wise weights in the LQ embeddings. Fixing all 13 weights to 1 yields the best LPIPS but lower NR-IQA scores in Tab. 2 and less perceptually favorable results in the third column of Fig. 5. In contrast, learnable weights adaptively reduce the influence of LQ at finer scales, allowing the model to better exploit generative priors and improve perceptual quality.

Ablation on $\mathcal{L}_{\text{SAF}}$. By employing $\mathcal{L}_{\text{SAF}}$, GPABC is driven to focus on the most critical correction behavior at each scale. Quantitative results in Tab. 2 demonstrate that $\mathcal{L}_{\text{SAF}}$ is beneficial for improving the perceptual quality of restored images by increasing no-reference metrics, accompanied by a slight improvement in LPIPS. The visual comparisons in Fig. 5 further corroborate this observation, where the details of the branches become noticeably smoothed when removing $\mathcal{L}_{\text{SAF}}$.

Table 2: Ablation studies on different components of GPABC on the DIV2K-Val benchmark.

Method	LPIPS↓	CLIPIQA↑	MUSIQ↑	TOPIQ↑
Fixed $\rho = 1$	**0.2562**	0.8632	73.00	0.7644
w/o $\mathcal{L}_{\text{SAF}}$	0.2782	0.7573	72.07	0.7415
w/o curriculum	0.2794	0.7974	72.77	0.7604
GPABC	0.2577	**0.8654**	**74.18**	**0.7838**

Necessity of curriculum learning. Full teacher-forcing training is adopted as the baseline for evaluating the curriculum learning strategy. By progressively mitigating exposure bias between training and autoregressive inference, curriculum learning achieves higher NR-IQA scores in Tab. 2 and alleviates the artifacts observed in Fig. 5 compared with full teacher-forcing.

6 Conclusion

We revisit image super-resolution as an LQ-anchored bitwise correction task. Our analysis reveals that LQ and HQ representations are highly similar at the bit level, while their differing bits follow structured patterns across scales, degradations, margins, and channels. These findings motivate GPABC, which corrects LQ bits toward their HQ counterparts through LQ-relative flip-mask prediction enhanced by a pretrained generative prior. A scale-adaptive asymmetric focal flip loss heavily penalizes false flips at coarse scales to preserve reliable LQ content, and missed flips at fine scales to synthesize high-frequency details. Extensive experiments demonstrate that GPABC outperforms state-of-the-art methods, achieving a superior balance between fidelity and perceptual quality.

AI USE STATEMENT

We used AI tools to aid and polish the writing of this manuscript. Their use was limited to improving grammar, clarity, conciseness, wording, and overall readability. AI tools were strictly limited to language refinement and were not used to develop research ideas, formulate the proposed method, or analyze results. All technical content, analysis, mathematical formulations, experimental results, and conclusions were independently developed by the authors. Every AI-assisted revision was carefully reviewed and edited to ensure technical accuracy and consistency with the authors' intended meaning. The authors take full responsibility for the content and integrity of this work.

REPRODUCIBILITY STATEMENT

In this manuscript, Sec. 4.2 describes GPABC, including shared-prefix bit computation, LQ-relative flip-mask prediction, the scale-adaptive asymmetric focal flip loss, and the curriculum learning strategy. Sec. 5.1 specifies the training and evaluation datasets, degradation pipeline, optimization settings, evaluation metrics, and experimental protocol. Appendix A provides further implementation details. We will release the source code and pretrained checkpoints after acceptance to facilitate the reproduction of our results.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1122–1131, 2017.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the International Conference on Computer Vision*, pp. 3086–3095, 2019.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:2567–2581, 2020.

Yushun Fang, Yuxiang Chen, Shibo Yin, Qiang Hu, Jiangchao Yao, Ya Zhang, Xiaoyun Zhang, and Yanfeng Wang. One-step diffusion transformer for controllable real-world image super-resolution. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 23440–23450, 2026.

Isma Hadji, Enrique Sanchez, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. Hierarchical image tokenization for multi-scale image super resolution. In *Proceedings of the International Conference on Machine Learning*, 2026.

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. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 15733–15744, 2025.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the International Conference on Computer Vision*, pp. 5128–5137, 2021.

Aro Kim, Myeongjin Jang, Chaewon Moon, Youngjin Shin, Jinwoo Jeong, and Sang-hyo Park. Fidesr: High-fidelity and detail-preserving one-step diffusion super-resolution. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 38270–38280, 2026.

Xiangtao Kong, Rongyuan Wu, Shuaizheng Liu, Lingchen Sun, and Lei Zhang. Nsarm: Next-scale autoregressive modeling for robust real-world image super-resolution. *arXiv preprint arXiv:2510.00820*, 2025.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 105–114, 2017.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Lsdir: A large scale dataset for image restoration. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1775–1787, 2023.

Cencen Liu, Dongyang Zhang, Wen Yin, Jielei Wang, Tianyu Li, Ji Guo, Wenbo Jiang, Guoqing Wang, and Guoming Lu. Alignvar: Towards globally consistent visual autoregression for image super-resolution. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 26286–26296, 2024.

Yunpeng Qu, Kun Yuan, Jinhua Hao, Kai Zhao, Qizhi Xie, Ming Sun, and Chao Zhou. Visual autoregressive modeling for image super-resolution. In *Proceedings of the International Conference on Machine Learning*, pp. 50926–50948, 2025.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 2333–2343, 2025.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in Neural Information Processing Systems*, 37:84839–84865, 2024.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2555–2563, 2023.

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*, 132:5929–5949, 2024.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshops*, pp. 63–79, 2018.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the International Conference on Computer Vision Workshops*, pp. 1905–1914, 2021.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.

Hongyang Wei, Shuaizheng Liu, Chun Yuan, and Lei Zhang. Perceive, understand and restore: Real-world image super-resolution with autoregressive multimodal generative models. In *Proceedings of the International Conference on Computer Vision*, pp. 18640–18650, 2025.

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 *Proceedings of the European Conference on Computer Vision*, pp. 101–117, 2020.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37: 92529–92553, 2024a.

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 25456–25467, 2024b.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1190–1199, 2022.

Fanghua Yu, Jinjin Gu, Zheyuan Li, Jinfan Hu, Xiangtao Kong, Xintao Wang, Jingwen He, Yu Qiao, and Chao Dong. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 25669–25680, 2024.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the International Conference on Computer Vision*, pp. 4771–4780, 2021.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Yu Zheng, Kai Zhang, Wei Zhu, Qingguo Liu, Xiantao Hu, Jun Li, and Jian Yang. Dvar: Dynamic visual autoregressive modeling for image super-resolution. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pp. 23378–23387, 2026.

Yixuan Zhu, Shilin Ma, Haolin Wang, Ao Li, Yanzhe Jing, Yansong Tang, Lei Chen, Jiwen Lu, and Jie Zhou. Varestorer: One-step var distillation for real-world image super-resolution. In *International Conference on Learning Representations*, 2026.

A Detailed Implementations

To exploit the generative prior of VAR, GPABC adopts Infinity-2B as its backbone, which is a next-scale prediction visual autoregressive model consisting of 32 transformer layers, an embedding dimension of 2048, and 16 attention heads. The model operates on a 32-dimensional BSQ latent space and autoregressively predicts bits across 13 resolution scales with grid sizes (1, 2, 4, 6, 8, 12, 16, 20, 24, 32, 40, 48, 64), targeting 1024×1024 output. We keep the entire Infinity backbone frozen, including VAE, text encoder and VAR transformer, and attach LoRA adapters with both the rank and scaling factor set to 8 to the cross-attention, self-attention, and feed-forward layers, yielding 224 adapters with 12.06M parameters. Two additional learnable linear projection layers, containing 0.14M parameters, embed the LQ BSQ bits and their normalized margins, with their outputs modulated by per-scale learnable weights. In total, only 12.2M parameters (0.5% of the backbone) are trainable. Training is conducted on 16 NVIDIA H20 GPUs for 22K iterations. As described in Sec. 5.1 of the main paper, we adopt a two-stage curriculum learning to mitigate exposure bias inherent in next-scale prediction. The first stage employs full teacher forcing for 2K iterations. During the remaining 20K iterations, predicted and HQ bits are mixed at their disagreement positions, with the probability p of using predicted bits increasing linearly from 0 to 1. During inference, following NSARM (Kong et al., 2025), we apply color alignment to the restored output using the LQ image as reference through wavelet-based color transfer and Adaptive Instance Normalization (AdaIN).

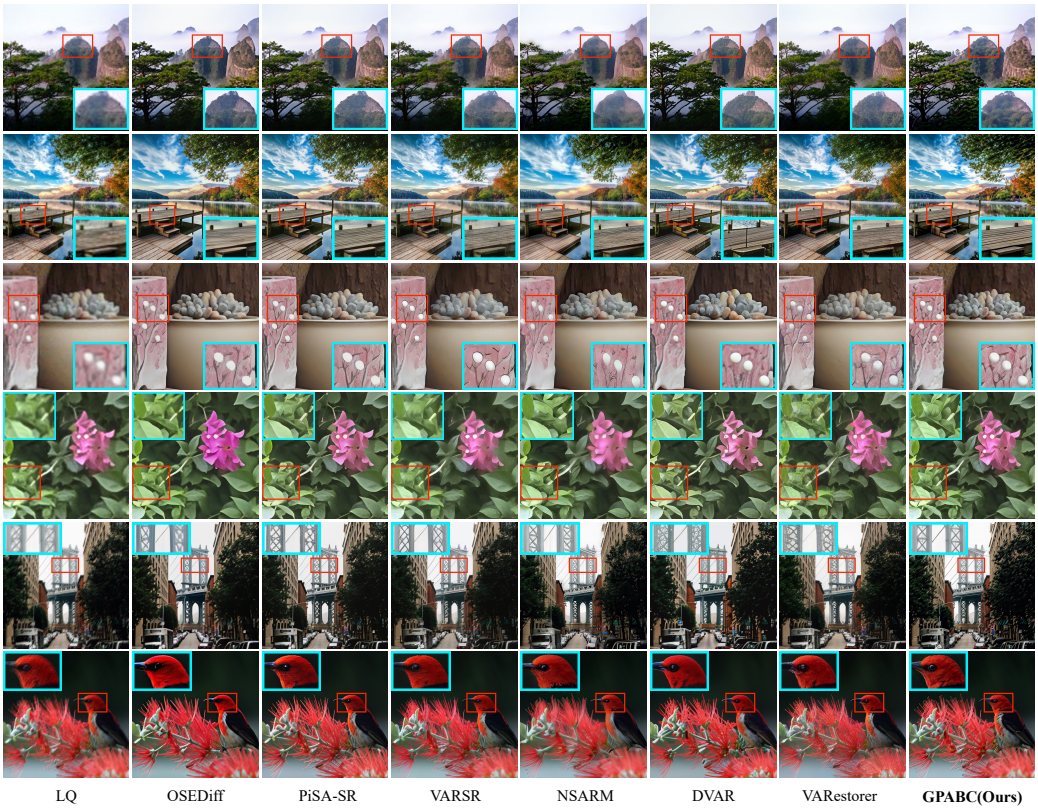


Figure A: Additional Qualitative Comparisons of the proposed GPABC and other SOTA methods. Please zoom in for a better view.

B Additional Qualitative Comparisons

More qualitative comparisons are provided in this section to further demonstrate the effectiveness of our framework. As shown in Fig. A, GPABC generates high-quality and highly realistic images while remaining faithful to the LQ images. Taking the first row as an example, VARSR, NSARM and DVAR

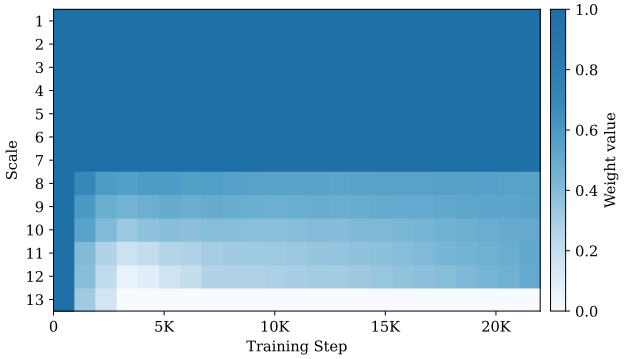


Figure B: Evolution of the scale-wise weights applied to the LQ embeddings across all 13 scales during training.

struggle to restore the vegetation morphology at the top of the mountain and introduce some artifacts. OSEDiff, PiSA-SR and VARestorer recover some structures of the trees, but their restorations remain limited in high-frequency details and intricate patterns. In contrast, the proposed GPABC retains fine-grained textures and subtle structural information, yielding sharper and more natural results. This highlights the strength of GPABC in balancing the perceptual quality and reconstruction fidelity.

C Ablation on encoding strategy

We examine the role of the shared prefix by comparing GPABC with an independent encoding variant. In the latter, the LQ and HQ images are quantized using their respective prefix codes from scales $0, \ldots, s-1$. Since code correction prediction at scale s is conditioned on all preceding scales, differences between the two prefixes produce distinct autoregressive contexts and quantization trajectories. Consequently, their bitwise disagreements may reflect not only image degradation but also context shifts accumulated at earlier scales. This ambiguity makes the relative flip mask a less stable correction target, as some predicted flips may compensate for inconsistent histories rather than degradation-induced discrepancies. In contrast, GPABC conditions the LQ and HQ branches on the same prefix. This shared context aligns their quantization trajectories and allows corresponding bits to be compared under a common autoregressive history. The resulting flip mask can therefore focus more directly on discrepancies associated with degradation. Quantitatively, independent encoding obtains 0.2734 LPIPS, 0.7802 CLIPIQA, 72.11 MUSIQ, and 0.7504 TOPIQ. Shared-prefix encoding improves these scores to 0.2577, 0.8654, 74.18, and 0.7838, respectively, demonstrating consistent gains in both reconstruction fidelity and perceptual quality. As shown in Fig. 5, independent encoding produces unnatural textures and local artifacts, whereas GPABC recovers more coherent structures and details. These results support the importance of maintaining a shared prefix context for reliable bitwise correction.

D Evolution of scale-wise weights across scales

Fig. B illustrates the evolution of the scale-wise weights applied to the LQ embeddings across all 13 scales during training. The weights at scales 1–7 remain at their initial value of 1, indicating that LQ information is fully utilized at coarse scales. In contrast, the weights at scales 8–12 decrease sharply during early training and subsequently stabilize or gradually recover to moderate values. This behavior suggests that LQ information remains useful at intermediate and fine scales but should be adaptively attenuated. Notably, the weight at scale 13, which is the finest scale, rapidly decreases toward 0 and remains suppressed thereafter, indicating that LQ information becomes unreliable at this scale and that the model primarily relies on generative priors to synthesize high-frequency details. These results demonstrate that the learned scale-wise weights adaptively balance LQ embeddings and generative priors across scales, preserving reliable coarse structures while reducing the influence of degraded LQ information at finer scales.

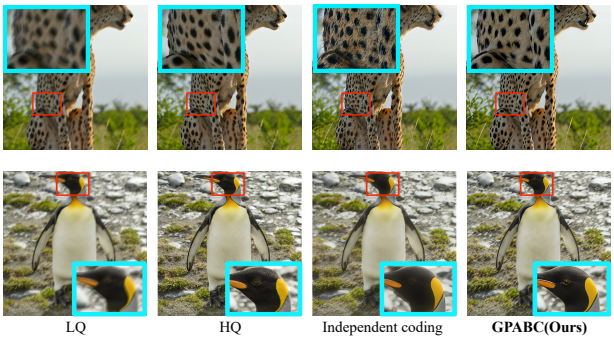


Figure C: Visualization for the ablation study on different encoding strategies. Please zoom in for a better view.

E INFERENCE TIME AND PARAMETER COMPARISON

Tab. A compares the inference time and parameter counts of GPABC with representative DM-based and VAR-based ISR methods. All methods are evaluated on an NVIDIA H20 GPU using an input image of size 128×128. GPABC requires 1.80 seconds per image, making it substantially faster than iterative DM-based methods such as StableSR, SeeSR, and SUPIR, which require 15.25, 4.03, and 8.60 seconds, respectively. PURE incurs substantially higher computational overhead, with its inference time reaching 240.26 seconds. Although one-step diffusion methods such as OSEDiff and PiSA-SR achieve faster inference, they still face challenges in balancing reconstruction fidelity and perceptual quality, as discussed in Sec. 5.2 of the main paper. In contrast, GPABC achieves a practical inference time of 1.80 seconds with only 12.2M trainable parameters while providing a more favorable balance between reconstruction fidelity and perceptual quality.

Table A: **Inference time and parameter comparison.** All methods are tested with an input image of size 128×128, and the inference time is measured on an NVIDIA H20 GPU.

	StableSR	SeeSR	SUPIR	OSEDiff	PiSA-SR	VARSR	VARestorer	PURE	DVAR	NSARM	GPABC
Inference Time (s)	15.25	4.03	8.60	0.15	0.09	0.49	0.40	240.26	1.04	1.20	1.80
# Total Param (M)	1,454	2,511	4,801	1,775	1,290	1,222	2,380	7,110	2,682	8,858	2,342
# Trainable Param (M)	149.9	749.9	1,298.2	8.5	8.1	1,113	27.3	7109.7	742.3	8761	12.2