PHADISR: PHASE-AWARE ONE-STEP DIFFUSION TRANSFORMER FOR REAL-WORLD IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world image super-resolution increasingly benefits from the strong generative priors of diffusion transformers (DiTs), yet one-step DiT super-resolution remains prone to periodic, grid-like artifacts. Using a translation-equivariance diagnostic, we trace these artifacts to *token-phase dependence*: prediction discrepancies in packed DiTs vary periodically with the packing stride, whereas a convolutional U-Net counterpart shows no such periodicity. Motivated by this diagnosis, we propose PhaDiSR, a phase-aware one-step diffusion transformer for real-world image super-resolution that addresses token-phase dependence through velocity correction and wavelet discrimination. At the prediction level, Polyphase Token Reassembly exposes the implicit token phases in the packed velocity prediction, models their local spatial relations, and reassembles a phase-aware correction before the endpoint update. To penalize the resulting artifacts, Token-Phase Wavelet Discriminator uses multi-scale, multi-shift Haar supervision to effectively capture diverse local token-phase groupings. To facilitate training, Self-Prior Projection uses the frozen initialization prior to construct projected training references. PhaDiSR eliminates periodic grid artifacts and achieves leading perceptual fidelity across standard benchmarks.

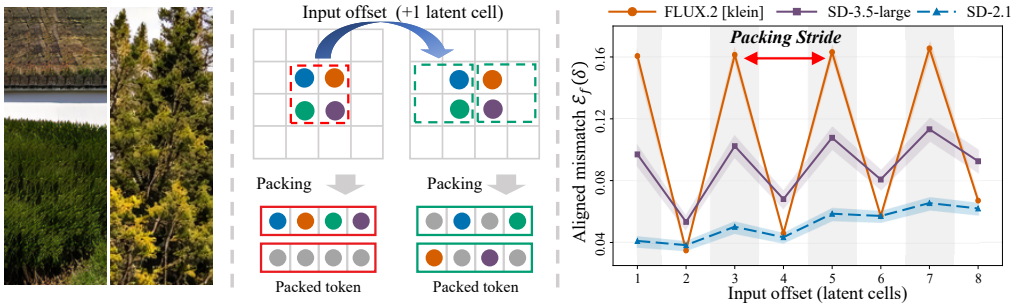
1 INTRODUCTION

Real-world image super-resolution (Real-ISR) aims to recover high-quality images from low-quality (LQ) inputs suffering from unknown, compound degradations. The task is ill-posed: the input constrains the overall scene structure, while high-frequency textures and fine details remain ambiguous. Pretrained diffusion models provide strong generative priors that help resolve this ambiguity, underpinning a prominent line of Real-ISR solutions (Wu et al., 2024a; Lin et al., 2025; Li et al., 2025).

Recently, diffusion backbones have increasingly transitioned from predominantly U-Net-based models to scalable diffusion transformers (DiTs), which offer stronger generative priors. Unlike convolutional U-Nets, however, DiTs process image representations through discrete tokens. While multi-step restoration requires iterative denoising, one-step models reduce inference cost by predicting the endpoint in a single forward pass. However, recent one-step DiT-based SR models have been observed to produce prominent periodic grid artifacts (Deng et al., 2026; Wang et al., 2026a), a phenomenon we also observe in Figure 1(a). The central challenge for one-step super-resolution is thus not only to exploit scalable DiT priors for one-step prediction, but also to maintain translation equivariance across token boundaries under the tokenized computation of the backbone.

In packed DiTs, *latent packing* groups neighboring latent positions into non-overlapping cells and maps each cell to one token. The periodic regularity of these grid-like artifacts suggests a direct connection to this fixed spatial organization. As illustrated in Figure 1(b), a token phase is the relative position within a latent packing cell; translating the same content can change this position and its packed-token assignment. We hypothesize that, if packing affects prediction consistency, the prediction should depend on how content is positioned relative to the packing grid, with a period determined by the packing stride. We test this hypothesis using a translation-equivariance diagnostic based on translation and inverse alignment. Figure 1(c) confirms this prediction: packed DiTs (FLUX.2 [klein] and SD-3.5-large) exhibit periodic variation in aligned prediction mismatch, while

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.



(a) Periodic artifacts (b) Token-phase dependence (c) Prediction mismatch vs. offset

Figure 1: **Diagnosing token-phase dependence in one-step DiT super-resolution.** (a) One-step DiT super-resolution produces periodic, grid-like artifacts. (b) Latent packing mechanism: a 2×2 packing cell groups four adjacent latent cells corresponding to distinct token phases into one packed token. Translating the input content (e.g., by $+1$ latent cell) alters its relative token phase and splits co-located contents across adjacent packing cells. (c) Translation-equivariance diagnostic: under inverse alignment, packed DiTs (FLUX.2 [klein], SD-3.5-large) exhibit periodic variation in prediction mismatch matching the packing stride ($s = 2$), whereas a U-Net counterpart (SD-2.1) shows no such periodicity.

a standard U-Net counterpart (SD-2.1) shows no comparable pattern. Crucially, the mismatch does not simply increase with displacement, but repeatedly rises and falls, exactly matching the packing stride. We term this sensitivity *token-phase dependence*. This observation exposes a fundamental mismatch between token-space computation and image-space spatial structure: **a reliable endpoint prediction should preserve the same local content regardless of its token phase.**

Motivated by this diagnosis, we introduce PhaDiSR, a phase-aware one-step diffusion transformer for real-world image super-resolution that addresses token-phase dependence through velocity correction and wavelet discrimination. Since token phase is implicit in packed channels, Polyphase Token Reassembly (PTR) explicitly exposes the token phases in the packed velocity prediction, models their local spatial relations, and reassembles a phase-aware correction before the endpoint update. To penalize the resulting artifacts in the image domain, Token-Phase Wavelet Discriminator (TPWD) provides multi-scale, multi-shift Haar supervision. Specifically, multi-scale decomposition aligns feature variations with the packing-cell scale, while multi-shift sampling dynamically evaluates diverse token-phase groupings, including those crossing cell boundaries. To facilitate model training, Self-Prior Projection (SPP) leverages the frozen generative prior to construct projected training references. Together, PTR and TPWD form a complementary defense against token-phase dependence, while SPP supplies the training targets. PhaDiSR eliminates periodic grid artifacts and achieves leading perceptual fidelity across standard benchmarks.

Our contributions are threefold: (i) we identify token-phase dependence in one-step DiT restoration, where a translation-equivariance diagnostic reveals that aligned prediction mismatch exhibits periodic variation matching the packing stride in packed DiTs, whereas a U-Net counterpart shows no comparable pattern; (ii) we propose PhaDiSR to address token-phase dependence, coupling PTR for velocity correction with TPWD for wavelet discrimination, where PTR coordinates token-phase relations within and across packing cells, while TPWD penalizes the associated artifacts through multi-scale, multi-shift Haar supervision; and (iii) we introduce SPP to construct projected training references using the frozen initialization prior, providing reliable training targets to eliminate periodic grid artifacts and achieve leading perceptual fidelity across standard benchmarks.

2 Related Work

2.1 Diffusion Models and Generative Priors

Diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020) have become a leading approach for learning complex image distributions. Latent diffusion models perform the generative process within a compressed latent space to enable efficient pretraining (Rombach et al., 2022; Podell et al., 2024). Recent years have witnessed a major architectural shift: the Diffusion

Transformer (DiT) uses a scalable transformer backbone in place of a convolutional U-Net (Peebles & Xie, 2023). At the training-objective level, flow matching learns velocity fields along prescribed probability paths (Lipman et al., 2023; Liu et al., 2023). SDXL (Podell et al., 2024) is a prominent U-Net-based latent diffusion model; newer text-to-image DiTs include SD3 (Esser et al., 2024), FLUX (Black Forest Labs, 2024), and Qwen-Image (Wu et al., 2025). By combining scalable transformer architectures with learned velocity fields, DiT-based diffusion models offer generative priors that scale with compute and model capacity.

2.2 Diffusion-Based Super-Resolution

Early U-Net-based diffusion models pioneered the use of generative priors for Real-ISR (Wang et al., 2024a; Lin et al., 2024; Yang et al., 2024; Wu et al., 2024b; Yu et al., 2024). To reduce latency, one-step methods adapt pretrained diffusion priors for restoration in a single forward pass (Wang et al., 2024b; Wu et al., 2024a; Dong et al., 2024; Lin et al., 2025; Xie et al., 2024; Sun et al., 2025; Chen et al., 2025). In parallel, the field is transitioning toward DiT backbones, which provide access to scalable generative priors. Recent advances have increasingly adopted DiT backbones for super-resolution, retaining multi-step iterative sampling (Duan et al., 2025; Ai et al., 2024) or supporting one-step predictions (Li et al., 2025; Kim et al., 2026; Fang et al., 2025; Gao et al., 2026; Wang et al., 2026b). Despite their efficiency, one-step DiT restoration can exhibit sensitivity to latent packing and periodic, grid-like artifacts (Wang et al., 2026a). Preliminary efforts have attempted to mitigate periodic artifacts (Deng et al., 2026; Wang et al., 2026a). PhaDiSR measures token-phase dependence and couples phase-aware velocity correction via PTR with multi-scale, multi-shift Haar supervision via TPWD, jointly targeting this sensitivity in both token representation and image space without altering the pretrained tokenization.

3 Method

PhaDiSR addresses the mismatch between token-space computation and image-space spatial structure in one-step super-resolution. Following the diagnostic in Section 3.1, Section 3.2 establishes the one-step formulation. To mitigate token-phase dependence, **Polyphase Token Reassembly (PTR)** models local spatial relations in the token velocity representation before the endpoint update (Section 3.3), while **Token-Phase Wavelet Discriminator (TPWD)** enforces multi-scale, multi-shift Haar supervision in the image domain (Section 3.4). Additionally, **Self-Prior Projection (SPP)** constructs projected training references via the frozen initialization prior (Section 3.5).

3.1 Diagnosing Token-Phase Dependence

Under non-overlapping latent packing with stride s, each packing cell groups s^2 adjacent latent locations into a single token, where a token phase denotes one of the s^2 relative positions within the cell (Figure 1(b)). While prevailing packed DiTs instantiate this mechanism with stride $s = 2$ (producing 2×2 cells), our diagnostic formulation applies generally to any stride s. Let Π and Π^{-1} denote packing and unpacking operators, and let f be the packed DiT evaluated at timestep $t_\star$. Its unpacked velocity prediction is $\mathcal{V}_f(u) = \Pi^{-1} f(\Pi(u), t_\star)$. For an integer displacement $\delta = (\delta_y, \delta_x)$, let T_δ denote spatial translation. Translating the noisy latent u and applying inverse alignment yields the translation-equivariance residual $\Delta_f(u, \delta) = M_\Omega \left[T_{-\delta} \mathcal{V}_f(T_\delta u) - \mathcal{V}_f(u)\right]$, where M_Ω restricts evaluation to the common valid interior Ω to exclude boundary effects.

$$\mathcal{E}_f(\delta) = \mathbb{E}_u \left[\frac{\|\Delta_f(u, \delta)\|_F^2}{C|\Omega|} \right], \tag{1}$$

where C is the latent channel dimension. The periodic variation matching the packing stride in Figure 1(c) confirms this systematic sensitivity as *token-phase dependence*.

This velocity mismatch matters directly for one-step restoration, whose endpoint map is

$$\mathcal{R}_f(u) = u - \sigma_\star \mathcal{V}_f(u), \tag{2}$$

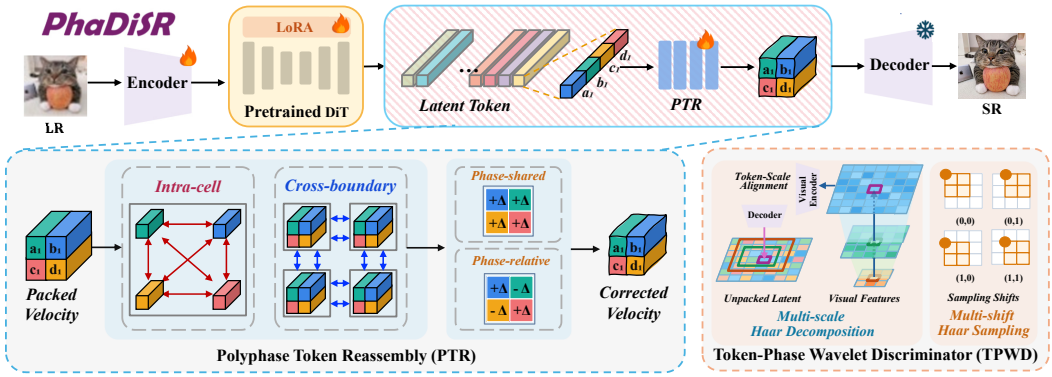


Figure 2: Overview of PhaDiSR. PTR coordinates the packed velocity prediction before the endpoint update, while TPWD penalizes periodic grid artifacts via multi-scale, multi-shift Haar supervision.

where $\sigma_\star$ is the scheduler coefficient at $t_\star$. On the common valid interior Ω, the discrepancy propagates directly to the latent endpoint:

$$M_\Omega\,[T_{-\delta}\mathcal{R}_f(T_\delta u) - \mathcal{R}_f(u)] = -\sigma_\star \Delta_f(u,\delta). \tag{3}$$

Eq. (3) shows that the endpoint residual equals the velocity residual scaled by $\sigma_\star$ in the valid interior. Thus, phase-dependent velocity errors can affect a one-step endpoint; their visibility in the image domain depends on their magnitude and the decoder response. This relation motivates correcting the velocity before the endpoint update, while the observed image artifacts are assessed empirically in Figure 1(a).

3.2 PhaDiSR Overview and One-Step Formulation

As shown in Figure 2, PhaDiSR integrates PTR into the one-step flow trajectory of a packed DiT before the endpoint update. The restoration backbone consists of a VAE encoder $\mathcal{E}_\theta$, packing and unpacking operators Π and Π^{-1}, a DiT velocity predictor f_θ, and a frozen VAE decoder $\mathcal{D}_0$. Given a low-quality input y, the encoder yields a latent feature $z_y \in \mathbb{R}^{2h\times 2w\times C}$, which Π packs into $N = hw$ tokens with dimension $4C$. Operating as a deterministic one-step mapping, the backbone evaluates the packed latent at timestep $t_\star$:

$$\begin{aligned} z_y = \mathcal{E}_\theta(y), \qquad P^0 = f_\theta(\Pi(z_y), t_\star)\,, \qquad P = \mathrm{PTR}_\psi(P^0), \\ \hat{z} = z_y - \sigma_\star \Pi^{-1}(P), \qquad \hat{x} = \mathcal{D}_0(\hat{z}) \equiv G_{\theta,\psi}(y). \end{aligned} \tag{4}$$

Here, P^0 and P denote the initial and corrected packed velocity predictions, respectively, with ψ collecting the trainable PTR parameters.

3.3 Polyphase Token Reassembly

Pretrained DiTs model global context across tokens, yet latent packing groups spatial positions into tokens and neglects local spatial relations within packing cells and across cell boundaries. To complement the backbone, PTR structures the packed velocity prediction $P^0 \in \mathbb{R}^{N\times 4C}$ $(N = hw)$ from Eq. (4) into an explicit polyphase representation by rearranging P^0 into four addressable phase components (detailed in Appendix E). PTR then executes two consecutive operations: *spatial feature coordination* and *orthogonal velocity correction*. First, spatial feature coordination aligns local relations through two stages: *intra-cell interaction* for the four phases within each packing cell and *cross-boundary interaction* for adjacent phases across cell boundaries. Second, orthogonal velocity correction projects the coordinated features into *phase-shared* and *phase-relative* components before writing them back to the packed prediction.

Intra-cell interaction. Each token phase attends to all four phases within the same packing cell, realigning co-located predictions before spatial writeback. The relative position bias B^{in} is parameterized by the pairwise phase displacement $(d_y^{q'} - d_y^q, d_x^{q'} - d_x^q)$, rendering the self-attention operation aware of local geometric offsets.

Cross-boundary interaction. This stage reconciles spatial continuity across packing cell boundaries, connecting adjacent predictions separated by the tokenization lattice. For a token phase at latent coordinate $p = (r, c) = (2i + d_y^q, 2j + d_x^q)$, the candidate neighborhood is

$$\mathcal{N}_{\mathrm{bd}}(r, c) = \big\{(r, c),\ (r + (-1)^{r+1}, c),\ (r, c + (-1)^{c+1})\big\}. \tag{5}$$

It comprises the phase itself and its two immediate neighbors across horizontal and vertical cell boundaries, with boundary-exceeding coordinates masked to yield a valid subset $\mathcal{N}_p^{\mathrm{valid}}$. The bias B_p^{bd} distinguishes directional adjacency relations.

Using multi-head attention with relative position bias B, the two interaction stages update the representations sequentially:

$$\begin{aligned} H_{i,j}^{(1)} &= H_{i,j}^{(0)} + W_o\Big(\mathrm{Attn}(\widetilde{Q}_{i,j}^{\mathrm{in}}, \widetilde{K}_{i,j}^{\mathrm{in}}, \widetilde{V}_{i,j}^{\mathrm{in}}; B^{\mathrm{in}})\Big), \\ H_p^{(2)} &= H_p^{(1)} + \mathrm{Attn}(q_p^{\mathrm{bd}}, k_{\mathcal{N}_p^{\mathrm{valid}}}^{\mathrm{bd}}, v_{\mathcal{N}_p^{\mathrm{valid}}}^{\mathrm{bd}}; B_p^{\mathrm{bd}}), \end{aligned} \tag{6}$$

where projection and normalization details follow Appendix E. A shared phase-wise GEGLU then produces the coordinated feature H. These sequential operations align local spatial relations within and across packing cells, preparing H for orthogonal velocity correction.

Phase-shared and phase-relative correction. After spatial feature coordination, PTR maps the refined representation $H_{i,j} \in \mathbb{R}^{4\times d}$ back to a velocity correction. A direct, unconstrained projection risks entangling collective token shifts with local phase adjustments, which can suppress valid image gradients. To prevent this coupling, PTR decomposes the correction into two orthogonal subspaces along the phase axis: a *phase-shared* (DC) component across the whole packing cell and a *phase-relative* (Ctr) component modeling intra-cell variations. Let $W_{\mathrm{dc}}, W_{\mathrm{ctr}} : \mathbb{R}^d \to \mathbb{R}^C$ denote affine projection heads applied across phases with shared parameters. With $\mathbf{1} \in \mathbb{R}^4$ the all-ones vector and $\mathbf{I}_4$ the identity matrix, we define orthogonal projectors:

$$\mathbf{P}_{\mathrm{dc}} = \tfrac{1}{4}\mathbf{1}\mathbf{1}^\top, \qquad \mathbf{P}_{\mathrm{ctr}} = \mathbf{I}_4 - \mathbf{P}_{\mathrm{dc}}, \qquad R_{i,j}^{\mathrm{dc}} = \mathbf{P}_{\mathrm{dc}} W_{\mathrm{dc}}(H_{i,j}), \qquad R_{i,j}^{\mathrm{ctr}} = \mathbf{P}_{\mathrm{ctr}} W_{\mathrm{ctr}}(H_{i,j}). \tag{7}$$

Here, the phase-shared correction $R_{i,j}^{\mathrm{dc}} \in \mathbb{R}^{4\times C}$ captures the collective DC shift of the packing cell, while the phase-relative correction $R_{i,j}^{\mathrm{ctr}} \in \mathbb{R}^{4\times C}$ captures zero-mean contrast variations across phases ($\sum_q R_{i,j,q}^{\mathrm{ctr}} = 0$). The orthogonality $\mathbf{P}_{\mathrm{dc}}\mathbf{P}_{\mathrm{ctr}} = \mathbf{0}$ guarantees strict functional separation: R^{dc} modulates collective token intensity, while R^{ctr} neutralizes packing-induced phase imbalance without suppressing structural gradients. We reassemble these polyphase corrections into the packed velocity prediction via:

$$P = P^0 + \rho^{-1}\big(\gamma_{\mathrm{dc}} R^{\mathrm{dc}} + \gamma_{\mathrm{ctr}} R^{\mathrm{ctr}}\big), \tag{8}$$

where ρ^{-1} maps the tensor back to the packed token layout. The learnable scalar gates γ_{dc} and γ_{ctr} are initialized to zero. This zero-gated residual design ensures that PTR operates as an exact identity mapping ($P = P^0$) at the onset of training, strictly preserving the base flow-matching trajectory before the phase-aware updates converge.

3.4 Token-Phase Wavelet Discriminator

While PTR coordinates phase-aware correction in the velocity representation, image-domain adversarial supervision is necessary to penalize artifacts that emerge after VAE decoding. Conventional image discriminators evaluate patches on a fixed coordinate grid, remaining blind to periodic phase discrepancies. To address this limitation, TPWD introduces two complementary geometric mechanisms: *multi-scale Haar decomposition* aligns discriminator feature scales with latent packing cells, and *multi-shift Haar sampling* perturbs evaluation windows to evaluate phase transitions across cell boundaries.

Multi-scale Haar decomposition. Discriminator features operate in the image domain, whereas token packing occurs in the latent space. To establish a direct geometric link between these two representations, we extract multi-level features $\{F_\ell(x)\}_{\ell=1}^{L}$ with spatial stride s_ℓ from a frozen visual encoder. Let r_{VAE} denote the decoder upsampling factor relative to the unpacked latent grid. Because a 2×2 packing cell occupies an image footprint of $2r_{\mathrm{VAE}} \times 2r_{\mathrm{VAE}}$, a 2×2 local patch

on feature level ℓ spans an image area of $2s_\ell \times 2s_\ell$. When the feature stride matches the decoder factor ($s_\ell = r_{\mathrm{VAE}}$), a 2×2 feature window precisely covers the physical footprint of one 2×2 latent packing cell. Applying a 2×2 Haar transform at this matched scale directly evaluates variations within individual packed tokens. Furthermore, varying the feature stride across levels naturally covers sub-token ($s_\ell < r_{\mathrm{VAE}}$), token-level ($s_\ell = r_{\mathrm{VAE}}$), and cross-token ($s_\ell > r_{\mathrm{VAE}}$) regimes.

Multi-shift Haar sampling. Under the matched stride $s_\ell = r_{\mathrm{VAE}}$, feature pixels and latent cells align on an identical spatial grid. Standard fixed-grid evaluation evaluates features within static cell boundaries, leaving it unable to detect artifacts that straddle adjacent cells. To disrupt this rigid grouping, we introduce a spatial grid shift $o = (o_y, o_x) \in \{0, 1\}^2$. Shifting the sampling window by a single feature pixel corresponds to a one-cell displacement on the latent grid. This shift forces the 2×2 Haar window to span across adjacent packing cells and evaluate alternative phase pairings. For each sampled shift o, the directional Haar transform decomposes the shifted feature map channel-wise via:

$$\widetilde{\mathcal{H}}_o^b(F_\ell)[i, j] = \sum_{a=0}^{1} \sum_{c=0}^{1} K^b[a, c]\, F_\ell[2i + a + o_y,\, 2j + c + o_x], \tag{9}$$

using four orthogonal 2×2 Haar kernels to decouple shared structures from signed directional variations:

$$K^{\mathrm{LL}} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad K^{\mathrm{LH}} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}, \quad K^{\mathrm{HL}} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, \quad K^{\mathrm{HH}} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \tag{10}$$

Following normalization and spatial resizing, we separate the shared component $\mathcal{H}_o^{\mathrm{LL}}$ from the concatenated directional variations $\mathcal{H}_o^{\mathrm{HF}} = \mathrm{Concat}(\mathcal{H}_o^{\mathrm{LH}}, \mathcal{H}_o^{\mathrm{HL}}, \mathcal{H}_o^{\mathrm{HH}})$. This orthogonal decomposition in the image domain directly mirrors the phase-shared and phase-relative formulation in PTR. We sample one shift per microbatch, evaluating dynamic spatial groupings across iterations without increasing training latency.

Wavelet discrimination. At each feature level, three discriminator heads evaluate the semantic features, shared LL structure, and signed HF variations:

$$\mathcal{D}_{\mathrm{TPWD}}^{\ell}(x; o) = \Big(D_\ell^{\mathrm{sem}}(F_\ell(x)), D_\ell^{\mathrm{LL}}(\mathcal{H}_o^{\mathrm{LL}}(F_\ell(x))), D_\ell^{\mathrm{HF}}(\mathcal{H}_o^{\mathrm{HF}}(F_\ell(x)))\Big), \qquad \ell = 1, \ldots, L. \tag{11}$$

The high-frequency heads D_ℓ^{HF} penalize phase discontinuities across dynamic cell partitions, D_ℓ^{LL} regulates local energy profiles, and the semantic head D_ℓ^{sem} preserves structural fidelity to the underlying image.

3.5 Self-Prior Projection

Supervised training of one-step restoration models with external high-quality targets often encounters a distributional discrepancy: arbitrary natural images do not necessarily align with the generative manifold of the pretrained backbone. To resolve this mismatch, **Self-Prior Projection (SPP)** constructs projected training references by mapping source images through the frozen initialization prior.

We freeze the pretrained backbone $(\mathcal{E}_0, f_0, \mathcal{D}_0)$ from Section 3.2, whose deterministic endpoint map at noise level σ is $\mathcal{R}_0(u, \sigma) = u - \sigma\Pi^{-1} f_0(\Pi(u), t(\sigma))$. Given an HQ source image x and Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, SPP perturbs the source latent at noise level σ and evaluates the endpoint:

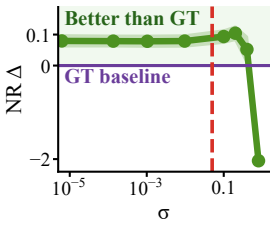


Figure 3: Perceptual quality across σ.

$$z_\sigma = (1 - \sigma)z + \sigma\epsilon, \qquad z = \mathcal{E}_0(x), \qquad z^+ = \mathcal{R}_0(z_\sigma, \sigma), \qquad x^+ = \mathcal{D}_0(z^+). \tag{12}$$

As derived in Appendix E, the latent displacement satisfies $\|z^+ - z\| = \mathcal{O}(\sigma)$. Crucially, mapping source images through this self-prior enhances their perceptual quality. As shown in Figure 3, the projection yields positive gains across no-reference perceptual metrics (NR $\Delta > 0$), confirming that the projected representations achieve higher visual quality while conforming to the native generative trajectory of the backbone. The resulting projected reference x^+ is cached offline, and its low-quality counterpart x_{LQ}^+ is synthesized by applying standard degradation operations to x^+.

3.6 Training Objective and Inference

We train PhaDiSR on the constructed pairs $(x^{+}_{\mathrm{LQ}}, x^{+})$ using three complementary objectives: low-frequency content preservation, wavelet-domain perceptual supervision, and multi-scale adversarial feedback. The content loss $\mathcal{L}_{\mathrm{blur}} = \mathrm{MSE}(\mathcal{G}(\hat{x}), \mathcal{G}(x^{+}))$ constrains structural profiles, where $\mathcal{G}$ denotes a Gaussian blur operator. The perceptual loss $\mathcal{L}_{\mathrm{perc}}$ measures feature distances across shifted Haar representations extracted by a frozen visual backbone.

For adversarial supervision, let $\mathcal{B}_{\phi}(x, a; o)$ denote the binary cross-entropy loss averaged across all logit maps from the TPWD heads parameterized by ϕ, given ground-truth label a and sampling shift o. With a label-smoothed target a_r, the generator and discriminator objectives are:

$$\begin{aligned} \mathcal{L}_G &= \lambda_{\mathrm{blur}}\mathcal{L}_{\mathrm{blur}} + \lambda_{\mathrm{perc}}\mathcal{L}_{\mathrm{perc}} + \lambda_{\mathrm{adv}}\mathcal{B}_{\phi}(\hat{x}, a_r; o), \\ \mathcal{L}_D &= \lambda_{\mathrm{adv}}\big[\mathcal{B}_{\phi}(\mathrm{stopgrad}(\hat{x}), 0; o) + \mathcal{B}_{\phi}(x^{+}, a_r; o)\big]. \end{aligned} \tag{13}$$

Optimization updates the LoRA parameters of the VAE encoder and DiT backbone alongside PTR, while the discriminator updates only its heads. At inference, PhaDiSR executes purely through the deterministic mapping in Eq. (4), requiring no discriminator evaluations.

4 Experiments

4.1 Experimental Settings

We train PhaDiSR on DIV2K (Agustsson & Timofte, 2017) and LSDIR (Li et al., 2023) using Real-ESRGAN degradations (Wang et al., 2021). Evaluation spans three paired benchmarks (RealSR (Cai et al., 2019), DrealSR (Wei et al., 2020), and ScreenSR (Wu et al., 2026)) and three unpaired real-world benchmarks, assessed via full-reference fidelity and no-reference perceptual metrics against U-Net and DiT baselines. We instantiate PhaDiSR on the FLUX.2 [klein] backbone at $t_{\star} = 500$ ($\sigma_{\star} \approx 0.5$), caching SPP references offline at perturbation $\sigma = 0.05$. Complete experimental settings, metric definitions, and baseline configurations are detailed in Appendix D.

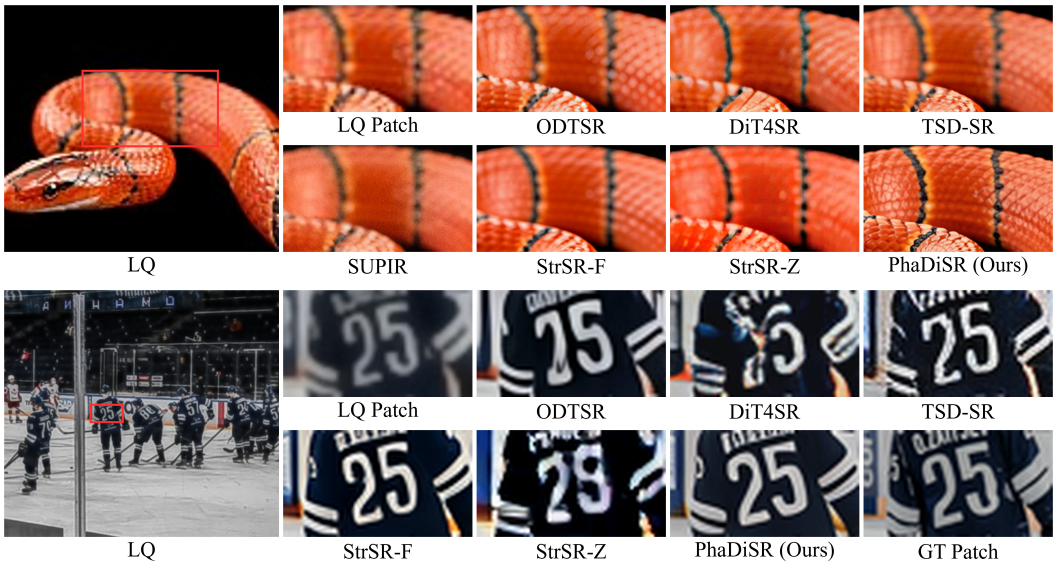


Figure 4: Qualitative comparison. The red box in the LQ image marks the region shown in the enlarged patches.

4.2 Comparison with the State of the Art

Quantitative Comparisons. Table 1 reports paired benchmark results; the unpaired benchmarks are reported together in Appendix A (Table 4). As shown in the tables, PhaDiSR establishes a

Table 1: Quantitative comparison on paired benchmarks. **Red** and blue indicate the best and second best results, respectively.

Set	Arch.	Method	Steps	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	A-FINE↓	MUSIQ↑	MANIQA↑	CLIPIQA+↑	TOPIQ-NR↑	LIQE↑	Q-Align↑	HyperIQA↑
RealSR	U-Net	ResShift	4	**25.66**	0.7360	0.3279	0.2475	52.72	56.90	0.5108	0.5181	0.4525	2.854	3.026	0.4604
		SeeSR	50	25.15	0.7210	0.3007	0.2224	24.06	69.81	0.6451	0.6911	**0.6891**	4.136	3.721	**0.6748**
		SUPIR	50	25.12	0.6523	0.3974	0.2560	50.14	58.82	0.5756	0.6144	0.5396	3.219	3.399	0.5449
		SinSR	1	**26.16**	**0.7373**	0.3070	0.2339	58.74	61.09	0.5423	0.5495	0.5334	3.194	3.189	0.5297
		OSEDiff	1	25.15	0.7341	0.2920	0.2128	20.94	69.10	0.6335	0.6966	0.6253	4.068	3.695	0.6260
		HYPIR	1	22.90	0.6859	0.3049	0.2271	11.25	66.21	0.6511	0.6805	0.5974	3.964	3.657	0.5993
	DiT	DiT4SR	40	23.65	0.6818	0.3101	0.2232	16.34	66.79	0.6453	0.6473	0.5882	3.869	3.278	0.6008
		TSD-SR	1	24.81	0.7172	0.2743	**0.2105**	18.05	**71.18**	0.6347	0.6883	0.6580	**4.267**	3.715	0.6435
		ODTSR	1	25.07	**0.7361**	**0.2395**	**0.1893**	15.19	68.25	**0.6622**	0.6520	0.5939	3.931	3.607	0.6140
		StrSR-F	1	21.93	0.7123	**0.2642**	0.2174	**9.795**	67.66	0.6383	0.6684	0.5873	3.925	3.613	0.5864
		StrSR-Z	1	21.12	0.6425	0.3248	0.2461	24.95	67.11	0.6027	**0.6981**	0.6561	3.814	**3.759**	0.6238
		PhaDiSR	1	24.30	0.7162	0.2999	0.2323	**5.079**	**71.28**	**0.6952**	**0.7097**	**0.7031**	**4.504**	**3.911**	**0.6985**
ScreenSR	U-Net	ResShift	4	24.55	**0.7636**	0.2557	0.1924	65.60	59.67	0.5276	0.5289	0.4997	2.833	2.645	0.5712
		SeeSR	50	22.45	0.6274	0.2233	**0.1371**	41.01	72.31	0.6900	0.7454	**0.7362**	4.642	**4.343**	**0.7310**
		SUPIR	50	24.02	0.7086	0.2126	0.1563	54.87	65.88	0.6425	0.6862	0.6135	4.052	3.802	0.6437
		SinSR	1	**24.76**	**0.7570**	0.2259	0.1779	62.46	66.25	0.5907	0.5916	0.6088	3.533	3.112	0.6534
		OSEDiff	1	22.12	0.6303	0.2303	0.1527	34.45	71.43	0.6809	0.7313	0.6282	4.609	4.014	0.6544
		HYPIR	1	**24.86**	0.7157	0.3982	0.2301	70.13	41.59	0.4932	0.5421	0.2887	2.063	2.338	0.3361
	DiT	DiT4SR	40	21.01	0.6053	0.2466	0.1602	40.42	71.35	0.6996	0.7005	0.6763	4.337	3.870	0.6847
		TSD-SR	1	21.60	0.6351	**0.1951**	0.1427	32.92	**74.12**	0.7032	0.7430	0.7170	**4.697**	4.149	0.7241
		ODTSR	1	21.99	0.6607	0.1980	**0.1331**	22.31	73.68	**0.7448**	**0.7483**	0.6962	4.668	4.145	0.7191
		StrSR-F	1	20.95	0.6818	0.2207	0.1640	**20.76**	72.59	0.7299	0.7465	0.6800	4.463	4.127	0.7039
		StrSR-Z	1	19.47	0.5751	0.2593	0.1671	47.93	71.98	0.6603	0.7364	0.7294	4.184	4.149	0.7082
		PhaDiSR	1	21.96	0.6856	**0.1922**	0.1439	**8.698**	**74.11**	**0.7623**	**0.7708**	**0.7486**	**4.811**	**4.380**	**0.7589**

Table 2: Generality across DiT foundations. Red and **blue** indicate the best and second best results, respectively.

Set	Backbone	Params (B)	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	MANIQA↑	CLIPIQA+↑	TOPIQ-NR↑	LIQE↑
ScreenSR	FLUX.1-dev	12	21.19	0.6032	0.2619	73.66	0.7294	0.7593	0.7326	4.773
	FLUX.2-Klein-9B	9	**22.54**	**0.7003**	**0.1786**	73.59	0.7557	0.7617	0.7305	4.780
	Z-Image-Turbo	6	22.05	0.6359	0.2388	71.92	0.6975	0.7109	0.6613	4.605
	SD3.5-Medium	2.5	21.06	0.6122	0.2566	73.42	0.7374	0.7619	0.7383	**4.790**
	Qwen-Image-2.1	7	**22.67**	0.6724	**0.1917**	73.74	0.7468	0.7408	0.7007	4.736
	Krea2-Raw	13	20.85	0.5953	0.2571	**74.66**	0.7604	0.7480	0.7253	4.783
	ERNIE-Image	8	21.61	0.6557	0.1987	74.00	**0.7681**	**0.7694**	**0.7470**	4.780
	PhaDiSR (Klein-4B)	4	21.96	**0.6856**	0.1922	**74.11**	**0.7623**	**0.7708**	**0.7486**	**4.811**

comprehensive advantage in no-reference perceptual metrics across diverse datasets. This structural superiority consistently extends to the unpaired RealLR200 dataset (Wu et al., 2024b), validating its robust perceptual performance in the wild. Comprehensive evaluations on additional datasets are provided in Appendix A.

Qualitative Comparisons. Figure 4 demonstrates the structural fidelity of PhaDiSR. It accurately reconstructs dense high-frequency textures, such as the snake scale lattice, where baselines yield smoothed or irregular artifacts. Furthermore, PhaDiSR strictly preserves geometric integrity under severe degradation, avoiding semantic hallucinations like the jersey number '25' distorted into '29' by StrSR-Z.

4.3 Generality Across DiT Foundations

To demonstrate that our proposed paradigm is model-agnostic and universally effective, we integrate the structural designs of PhaDiSR into a diverse spectrum of leading DiT foundations (Table 2). The cross-backbone benchmark reveals that our phase-aware correction mechanism successfully optimizes all evaluated architectures. This confirms that mitigating latent packing artifacts universally elevates one-step SR performance regardless of the specific foundation model. Overall, this extensive evaluation establishes PhaDiSR not as a singular specialized network, but as a highly generalizable and effective solution for the broader DiT family. Further configurations are detailed in Appendix A.

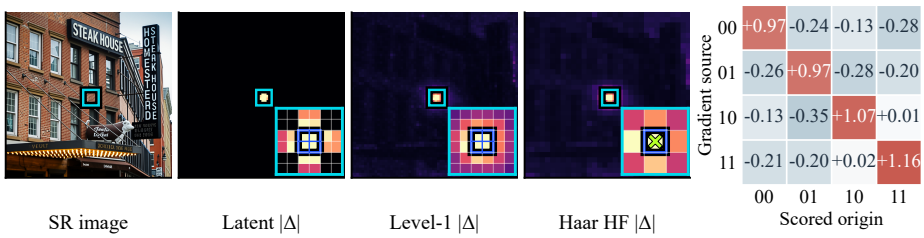


Figure 5: TPWD response to a packed token intervention. Left: reference image and difference maps across latent, feature, and Haar HF domains for a highlighted 2×2 cell. Right: HF head score responses under gradient updates evaluated across different sampling shifts.

Table 3: Ablation study of PhaDiSR.

Variant	PSNR↑	LPIPS↓	MUSIQ↑	MANIQA↑	CLIPIQA+↑
w/o PTR	24.02	**0.2910**	68.61	0.6723	0.6735
w/o TPWD	23.75	0.3176	**70.14**	**0.6819**	**0.6954**
w/o SPP	**24.36**	**0.2905**	69.87	0.6771	0.6862
PhaDiSR	**24.30**	0.2999	**71.28**	**0.6952**	**0.7097**

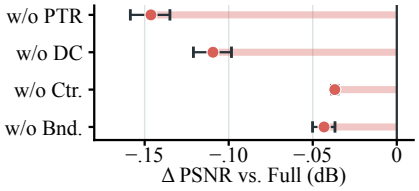


Figure 6: Ablation study for PTR.

4.4 Ablation Studies

Component Effectiveness. Table 3 validates the essential role of each proposed module. The full PhaDiSR configuration achieves the optimal visual fidelity, yielding the highest scores across perceptual metrics such as MUSIQ and CLIPIQA+. Omitting the SPP module biases the output toward pixel-level smoothness, thereby marginally improving PSNR but severely degrading high-frequency perceptual quality. Removing the PTR module disrupts phase-aware velocity correction, leading to a substantial drop in overall perception. Similarly, excluding the TPWD module removes critical multi-shift image-space supervision, further diminishing visual performance. These consistent degradations confirm that coupling velocity-space correction with image-space discrimination is indispensable for mitigating grid artifacts.

PTR Internal Coordination. In Figure 6, we intervene on the internal routing of PTR during inference to dissect its spatial mechanisms. Bypassing PTR entirely causes the most severe performance degradation. Furthermore, selectively disabling the phase-shared correction (w/o DC), phase-relative correction (w/o Ctr.), or cross-boundary interaction (w/o Bnd.) uniformly degrades restoration fidelity. This confirms that both orthogonal phase decomposition and cross-boundary continuity are essential to constructing the high-fidelity endpoint.

TPWD Local Correspondence and Feedback. Figure 5 demonstrates the spatial precision and origin-conditioned feedback of TPWD. As shown on the left, when intervening on a specific token phase in the packed prediction, the maximal high-frequency variation maps directly to the exact corresponding Haar window at Level 1. This confirms that TPWD targets precise local token phases rather than applying a generic image-wide penalty. Furthermore, the right panel illustrates the score matrix when evaluating a fixed cell across four distinct Haar grid shifts. The strongly positive diagonal entries establish that taking a gradient step for a specific shift strictly maximizes its corresponding local score, while cross-shift responses remain predominantly negative. This proves that multi-shift sampling successfully provides diverse, complementary supervision to dismantle rigid cell boundaries.

5 Conclusion

In this paper, we identify token-phase dependence as the fundamental structural bottleneck responsible for periodic grid artifacts in one-step diffusion transformer super-resolution. We propose PhaDiSR, coupling Polyphase Token Reassembly (PTR) for explicit velocity-space correction with Token-Phase Wavelet Discriminator (TPWD) for dynamic, multi-shift image-space supervision. Self-Prior Projection (SPP) constructs reliable training references using the frozen initialization prior. Extensive benchmark evaluations confirm that PhaDiSR establishes leading perceptual fidelity. Crucially, our results demonstrate that this phase-aware correction mechanism generalizes effectively across diverse DiT foundations, proving that dedicated structural optimization is a more robust and efficient path to high-quality one-step restoration than brute-force parameter scaling.

AI USE STATEMENT

In preparing this manuscript, generative AI tools were used solely for language editing, grammar proofreading, and LaTeX formatting. All architectural designs, theoretical formulations, code implementations, experimental evaluations, and scientific interpretations were independently conceived, conducted, and verified by the authors.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. Our research focuses on real-world image super-resolution without involving human subjects, demographic profiling, or private biometric collection. All datasets used for training and evaluation (DIV2K, LSDIR, and public test benchmarks) are publicly accessible and used strictly for academic research under their respective licenses. As with all generative restoration models, care should be taken to prevent generating misleading photographic records.

REPRODUCIBILITY STATEMENT

All mathematical formulations, architectural designs, objective functions, and training configurations are detailed in Section 3 and Appendices D–E. Complete evaluation protocols, metric definitions, and public benchmark splits are documented in Appendix D. The full source code, model checkpoints, and evaluation scripts will be made publicly available upon paper acceptance.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *CVPR Workshops*, 2017.

Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Xiaotian Han, Zhengyu Chen, Quanzeng You, and Hongxia Yang. Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. In *Advances in Neural Information Processing Systems*, 2024.

Black Forest Labs. Flux. `https://blackforestlabs.ai`, 2024.

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 *ICCV*, 2019.

Bin Chen, Gehui Li, Rongyuan Wu, Xindong Zhang, Jie Chen, Jian Zhang, and Lei Zhang. Adversarial diffusion compression for real-world image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Chengyan Deng, Pengbin Yu, Zhentao Chen, Wei Shen, Kai Zhang, Meng Li, Lunxi Yuan, Xue Zhou, and Li Yu. Op4ksr: One-step patch-free 4k super-resolution with periodic artifact suppression. *arXiv preprint arXiv:2605.13457*, 2026.

Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. *arXiv preprint arXiv:2411.18263*, 2024.

Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S. Ren, Chunle Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. *arXiv preprint arXiv:2503.23580*, 2025.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

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. *arXiv preprint arXiv:2511.17138*, 2025.

Xiang Gao, Chenxin Zhu, Yushun Fang, Qiang Hu, and Xiaoyun Zhang. Teasr: Training-efficient any-step diffusion transformer for real-world image super-resolution. *arXiv preprint arXiv:2606.16188*, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

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. *arXiv preprint arXiv:2603.02692*, 2026.

Jianze Li, Jiezhang Cao, Yong Guo, Wenbo Li, and Yulun Zhang. One diffusion step to real-world super-resolution via flow trajectory distillation. *arXiv preprint arXiv:2502.01993*, 2025.

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 *CVPR Workshops*, 2023.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior. In *European Conference on Computer Vision*, 2024.

Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S. Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *arXiv preprint arXiv:2507.20590*, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Levi. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, 2019.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 2024a. doi: 10.1007/s11263-024-02168-7.

Jingkai Wang, Yixin Tang, Jue Gong, Jiatong Li, Shu Li, Libo Liu, Jianliang Lan, Yutong Liu, and Yulun Zhang. Spectral and trajectory regularization for diffusion transformer super-resolution. *arXiv preprint arXiv:2603.06275*, 2026a.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *ICCV Workshops*, 2021.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. Sinsr: Diffusion-based image super-resolution in a single step. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Zihan Wang, Xudong Huang, Junbo Qiao, Wei Li, Jie Hu, Xinghao Chen, and Shaohui Lin. Allosr^2: Rectifying one-step super-resolution to stay real via allomorphic generative flows. *arXiv preprint arXiv:2604.19238*, 2026b.

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*, 2020.

Chenfei Wu et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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

Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, Xiangtao Kong, Jixin Zhao, Shihao Wang, and Lei Zhang. Vosr: A vision-only generative model for image super-resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16311–16321, 2026.

Rui Xie, Chen Zhao, Kai Zhang, Zhenyu Zhang, Jun Zhou, Jian Yang, and Ying Tai. Addsr: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. In *Advances in Neural Information Processing Systems*, 2024.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *European Conference on Computer Vision*, 2024.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

A Additional Quantitative Results

Extended Dataset Evaluations. Table 5 details the quantitative evaluation on the paired DrealSR benchmark, where PhaDiSR establishes a consistent advantage in structural fidelity and perceptual metrics over existing single-step baselines. Table 4 extends this evaluation to three unpaired real-world datasets (RealLR200, RealLQ250, and RealPhoto60). Across these diverse unconstrained degradations, PhaDiSR achieves superior reference-free perceptual scores, validating the robust generalization of phase-aware velocity correction for the real-world restoration.

Cross Architecture Generalization. Tables 6 and 7 report the comprehensive cross architecture evaluations, integrating the proposed phase-aware correction mechanism into a diverse spectrum of leading diffusion transformer foundations. The quantitative gains confirm that mitigating latent packing dependencies elevates restoration performance regardless of the underlying model scale. This establishes the dual domain correction paradigm as a fundamental architectural optimization applicable across distinct foundation variants.

Table 4: Quantitative evaluation on unpaired RealLR200, RealLQ250, and RealPhoto60 benchmarks. Red/blue: best/second per dataset.

Set	Arch.	Method	Steps	MUSIQ↑	MANIQA↑	CLIPIQA+↑	TOPIQ-NR↑	LIQE↑	Q-Align↑	HyperIQA↑
RealLR200	U-Net	ResShift	4	59.86	0.5260	0.5773	0.4971	3.054	3.798	0.5146
		SeeSR	50	69.60	0.6124	0.7069	**0.6630**	4.039	4.237	**0.6562**
		SUPIR	50	61.31	0.5880	0.6306	0.5331	3.346	4.109	0.5496
		SinSR	1	65.09	0.5505	0.6246	0.5990	3.598	3.939	0.5923
		OSEDiff	1	69.42	0.6053	0.7054	0.6048	4.052	4.259	0.6110
		HYPIR	1	69.36	0.6387	**0.7309**	0.6529	4.142	4.271	0.6439
	DiT	DiT4SR	40	70.20	0.6323	0.6985	0.6300	4.226	4.123	0.6316
		TSD-SR	1	**72.06**	0.6151	0.7172	0.6572	**4.284**	4.230	0.6440
		ODTSR	1	69.96	**0.6482**	0.6880	0.6059	4.043	**4.299**	0.6254
		StrSR-F	1	68.86	0.6324	0.7003	0.5970	4.010	4.264	0.6020
		StrSR-Z	1	68.77	0.5901	0.7044	0.6446	3.827	4.249	0.6117
		PhaDiSR	1	**72.41**	**0.6801**	**0.7506**	**0.7009**	**4.531**	**4.403**	**0.6929**
RealLQ250	U-Net	ResShift	4	59.51	0.5005	0.5532	0.4693	2.767	3.632	0.4805
		SeeSR	50	70.38	0.5927	0.7033	**0.6635**	3.979	4.143	**0.6442**
		SUPIR	50	62.34	0.5792	0.6170	0.5106	3.272	4.095	0.5251
		SinSR	1	65.30	0.5261	0.6041	0.5780	3.354	3.754	0.5615
		OSEDiff	1	69.55	0.5782	0.6981	0.6072	3.904	4.249	0.6002
		HYPIR	1	69.75	0.6075	**0.7195**	0.6351	4.069	4.227	0.6161
	DiT	DiT4SR	40	71.16	**0.6161**	0.6909	0.6228	4.225	4.121	0.6189
		TSD-SR	1	**73.22**	0.5924	0.7131	0.6463	**4.319**	4.177	0.6257
		ODTSR	1	69.63	0.6161	0.6624	0.5676	3.811	4.220	0.5898
		StrSR-F	1	69.89	0.6154	0.6964	0.5811	4.014	**4.287**	0.5827
		StrSR-Z	1	69.32	0.5775	0.6920	0.6332	3.804	4.279	0.5937
		PhaDiSR	1	**73.12**	**0.6613**	**0.7450**	**0.6925**	**4.541**	**4.367**	**0.6757**
RealPhoto60	U-Net	ResShift	4	58.31	0.4707	0.5493	0.4970	2.655	3.073	0.4953
		SeeSR	50	71.75	0.6077	0.7271	**0.6824**	4.123	4.320	**0.6471**
		SUPIR	50	68.71	0.6104	0.6808	0.5854	3.937	4.488	0.5910
		SinSR	1	62.82	0.4573	0.6026	0.5895	3.018	3.405	0.5663
		OSEDiff	1	70.46	0.5891	0.7167	0.6341	4.050	4.437	0.6180
		HYPIR	1	71.38	0.6112	**0.7442**	0.6670	4.313	4.392	0.6395
	DiT	DiT4SR	40	73.05	**0.6243**	0.7061	0.6460	4.482	4.210	0.6395
		TSD-SR	1	**74.91**	0.6101	0.7356	0.6718	**4.555**	4.413	0.6428
		ODTSR	1	70.31	0.6193	0.6814	0.5702	3.998	4.333	0.5956
		StrSR-F	1	71.40	0.6215	0.7204	0.5968	4.279	**4.527**	0.5949
		StrSR-Z	1	71.00	0.5957	0.7217	0.6680	4.156	4.477	0.6223
		PhaDiSR	1	**74.14**	**0.6642**	**0.7567**	**0.7046**	**4.678**	**4.538**	**0.6782**

Table 5: Quantitative comparison on paired benchmarks.

Set	Arch.	Method	Steps	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	A-FINE↓	MUSIQ↑	MANIQA↑	CLIPIQA+↑	TOPIQ-NR↑	LIQE↑	Q-Align↑	HyperIQA↑
DrealSR	U-Net	ResShift	4	27.05	0.7404	0.3871	0.2632	57.58	51.25	0.4666	0.5052	0.4530	2.675	2.861	0.4609
		SeeSR	50	28.07	0.7684	0.3174	0.2315	22.78	65.08	0.6052	0.6793	**0.6575**	4.127	3.588	**0.6641**
		SUPIR	50	26.86	0.6464	0.4593	0.2828	54.91	52.74	0.5242	0.5861	0.5067	2.963	3.189	0.5408
		SinSR	1	**28.21**	0.7546	0.3492	0.2484	58.53	57.87	0.5050	0.5678	0.5352	3.307	3.206	0.5360
		OSEDiff	1	27.92	**0.7835**	0.2967	0.2162	21.40	64.69	0.5898	**0.6822**	0.6002	3.940	3.540	0.6080
		HYPIR	1	26.07	0.7285	0.3282	0.2307	16.14	61.39	0.6024	0.6457	0.5688	3.761	3.542	0.5765
	DiT	DiT4SR	40	26.44	0.7148	0.3353	0.2325	15.34	62.91	0.6060	0.6293	0.5523	3.785	3.251	0.5731
		TSD-SR	1	27.77	0.7559	0.2966	**0.2136**	17.35	**66.61**	0.5874	0.6756	0.6181	**4.129**	**3.604**	0.6123
		ODTSR	1	**28.16**	**0.7740**	**0.2591**	**0.1926**	20.29	62.86	**0.6225**	0.6310	0.5597	3.574	3.462	0.5822
		StrSR-F	1	24.28	0.7468	**0.2815**	0.2211	**14.17**	61.72	0.5896	0.6370	0.5259	3.588	3.385	0.5336
		StrSR-Z	1	23.47	0.6877	0.3447	0.2454	19.53	62.68	0.5650	0.6506	0.6004	3.607	3.451	0.5827
		PhaDiSR	1	27.03	0.7491	0.3258	0.2388	**6.637**	**68.12**	**0.6587**	**0.6972**	**0.6745**	**4.374**	**3.794**	**0.6759**

Table 6: Generality across DiT foundations.

Set	Backbone	Params (B)	PSNR↑	SSIM↑	LPIPS↓	MUSIQ↑	MANIQA↑	CLIPIQA+↑	TOPIQ-NR↑	LIQE↑
RealSR	FLUX.1-dev	12	24.03	0.7005	0.3156	70.57	0.6728	0.6973	0.6758	4.433
	FLUX.2-Klein-9B	9	**25.04**	**0.7349**	**0.2840**	70.89	**0.6887**	0.6957	0.6849	4.361
	Z-Image-Turbo	6	24.76	0.7110	0.3012	67.55	0.6337	0.6605	0.6097	3.984
	SD3.5-Medium	2.5	24.04	0.7074	0.3162	70.44	0.6703	**0.7056**	**0.6908**	4.451
	Qwen-Image-2.1	7	**25.30**	**0.7320**	**0.2698**	70.66	0.6790	0.6680	0.6603	4.230
	Krea2-Raw	13	24.09	0.6961	0.3089	**71.37**	0.6823	0.6856	0.6838	**4.473**
	ERNIE-Image	8	24.66	0.7199	0.2861	70.39	0.6848	0.6869	0.6798	4.329
	PhaDiSR (Klein-4B)	4	24.30	0.7162	0.2999	**71.28**	**0.6952**	**0.7097**	**0.7031**	**4.504**
DrealSR	FLUX.1-dev	12	27.09	0.7618	0.3275	66.94	0.6311	0.6748	0.6363	4.302
	FLUX.2-Klein-9B	9	27.72	0.7734	**0.3000**	67.38	0.6508	0.6811	0.6555	4.269
	Z-Image-Turbo	6	**27.85**	**0.7772**	0.3127	63.73	0.5944	0.6502	0.5893	3.738
	SD3.5-Medium	2.5	27.08	0.7644	0.3218	66.50	0.6256	**0.6828**	0.6483	4.319
	Qwen-Image-2.1	7	**28.09**	**0.7766**	**0.2888**	67.08	0.6431	0.6547	0.6380	4.161
	Krea2-Raw	13	27.28	0.7632	0.3235	**68.43**	0.6405	0.6596	0.6611	**4.375**
	ERNIE-Image	8	27.56	0.7629	0.3114	67.41	**0.6538**	0.6793	**0.6611**	4.252
	PhaDiSR (Klein-4B)	4	27.03	0.7491	0.3258	**68.12**	**0.6587**	**0.6972**	**0.6745**	**4.374**

Table 7: Generality across DiT foundations.

Backbone	Params (B)	MUSIQ↑	MANIQA↑	CLIPIQA+↑	TOPIQ-NR↑	LIQE↑	Q-Align↑	HyperIQA↑
FLUX.1-dev	12	73.19	0.6360	0.7349	0.6678	4.537	**4.590**	0.6452
FLUX.2-Klein-9B	9	73.88	0.6589	**0.7516**	0.7036	4.646	4.523	0.6715
Z-Image-Turbo	6	70.56	0.6012	0.7039	0.6158	4.171	4.546	0.5953
SD3.5-Medium	2.5	73.88	0.6362	0.7444	0.6984	**4.687**	4.454	0.6672
Qwen-Image-2.1	7	**74.69**	0.6618	0.7287	0.6730	4.582	4.579	0.6688
Krea2-Raw	13	**74.63**	0.6551	0.7270	0.6767	4.676	**4.596**	0.6675
ERNIE-Image	8	73.95	**0.6636**	0.7505	**0.7064**	4.648	4.487	**0.6759**
PhaDiSR (Klein-4B)	4	74.14	**0.6642**	**0.7567**	**0.7046**	**4.678**	4.538	**0.6782**

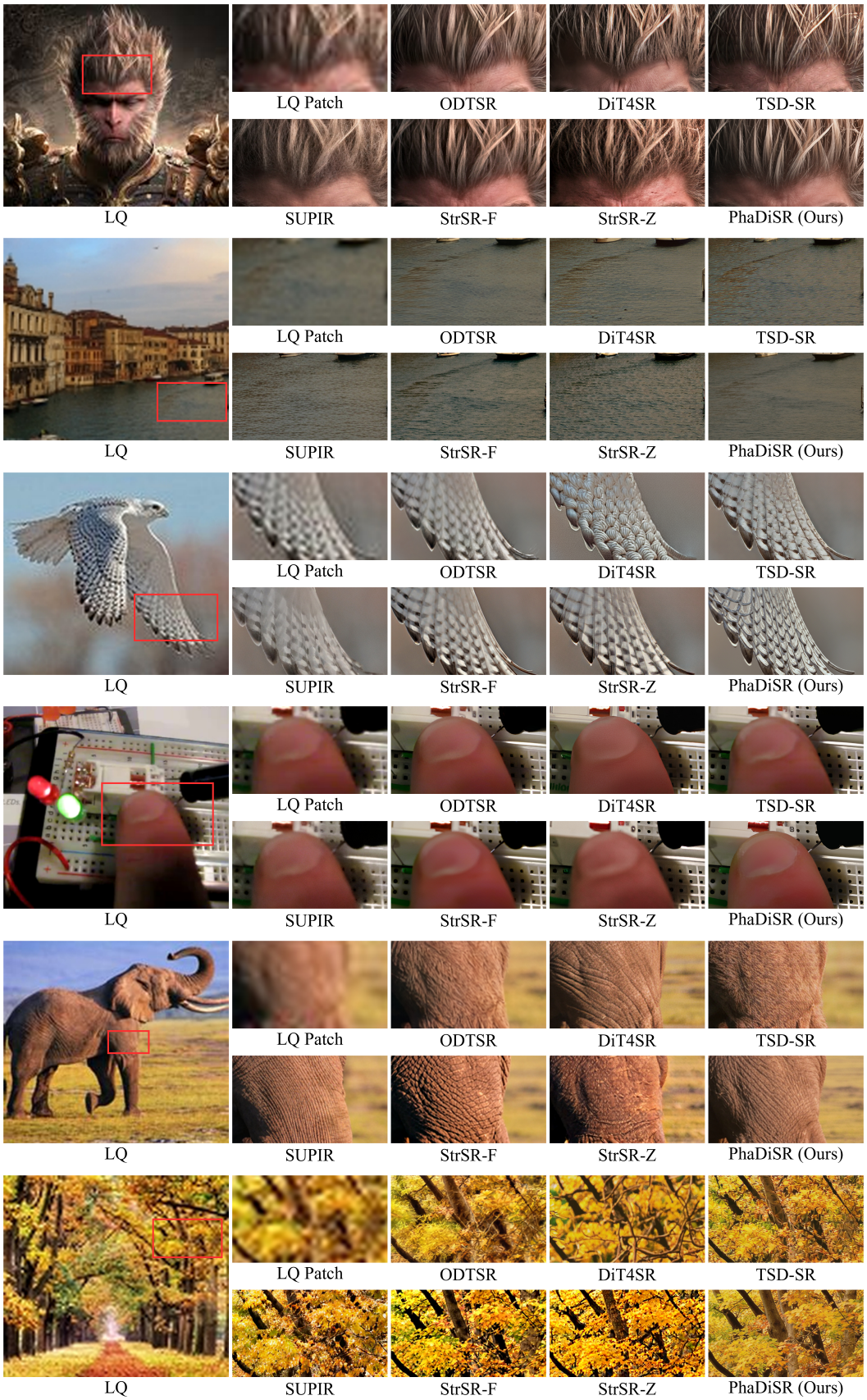


Figure 7: Qualitative comparison. The red box in the LQ image marks the region shown in the enlarged patches.

B More Visual Comparisons

Figures 7 provide extended visual comparisons across diverse scenes. Standard U-Net and baseline diffusion transformer architectures frequently exhibit periodic grid artifacts and structural blurring. In contrast, PhaDiSR effectively mitigates token phase dependencies to reconstruct accurate high frequency details and maintain rigid geometric boundaries.

C Diagnostic and Mechanism Visualizations

Mechanism visualizations evaluate a converged generator state. For a 512×512 image, the Level 1 feature map occupies a 64×64 spatial grid. Consequently, the critically sampled 2×2 Haar operator yields a 32×32 window lattice, strictly aligning with the spatial dimensions of the DiT packed token representation. This geometric congruence establishes the foundation for spatial correspondence tests. Similarly, PTR evaluations measure deterministic inference interventions on this converged backbone, complementing the independent component ablation studies.

C.1 TPWD: Origin Dependent Grouping and Returned Feedback

Operator geometry and structural feedback. The diagnostic isolates the Level 1 feature cell region $[32{:}34, 32{:}34]$ and systematically permutes the Haar origin. In Figure 8, the first row establishes this specific region at the center of four local coordinate systems. Magenta boundaries outline the explicit 2×2 Haar groups intersecting the defined target.

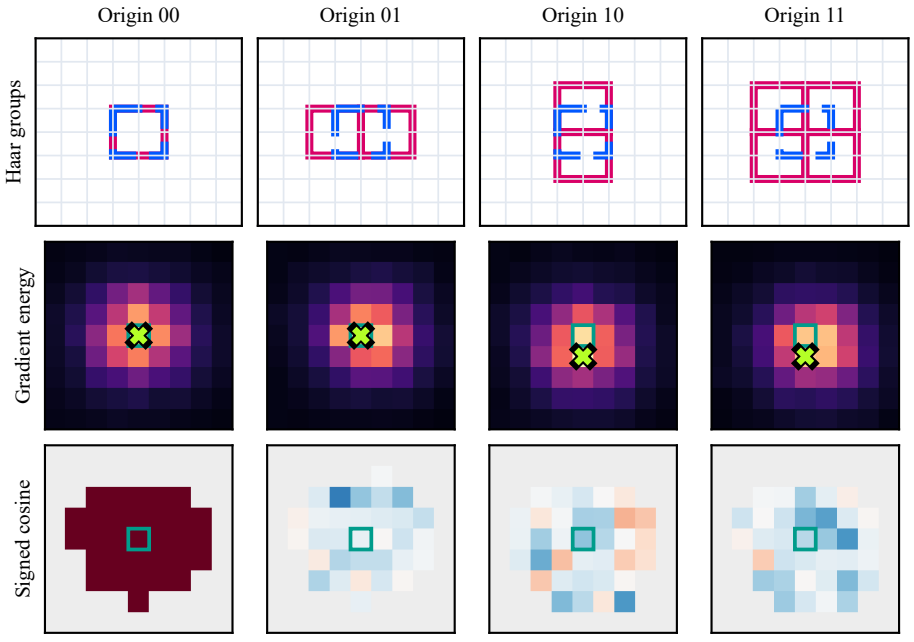


Figure 8: Evaluations of a fixed Level 1 target organized by Haar origin variables (columns). Top: the dashed blue frame indicates the fixed 2×2 feature region; magenta lines define intersecting Haar windows. Middle: normalized gradient energy propagated to packed tokens. Bottom: signed channel cosine evaluations against origin 00.

Evaluating a representative structured scene (`ScreenSR_003`), the diagnostic differentiates the identified Level 1 learned high frequency cell $(16, 16)$ with respect to the comprehensive $32 \times 32 \times 128$ packed prediction tensor. The marked target token aggregates approximately $15\%, 15\%, 12\%, 11\%$ of the propagated energy under origins $00, 01, 10, 11$, confirming robust, origin conditioned local feedback delivery across structural permutations.

C.2 TPWD: Critical Haar Window Correspondence

Predicted versus measured token responses. For a packed token at spatial index (i, j), phase (p_y, p_x), and Haar origin (d_y, d_x), the theoretical critical window index evaluates strictly as $(\lfloor (2i + p_y - d_y)/2 \rfloor, \lfloor (2j + p_x - d_x)/2 \rfloor)$. The diagnostic intervenes on a single phase by substituting it with its same phase neighborhood mean, subsequently measuring the channel RMS variation of the signed critical Haar coefficients immediately following the transform. Figure 9 illustrates this structural response. The initial evaluation perturbs the entire packed token, whereas subsequent configurations isolate interventions exclusively to phase 0 and phase 3.

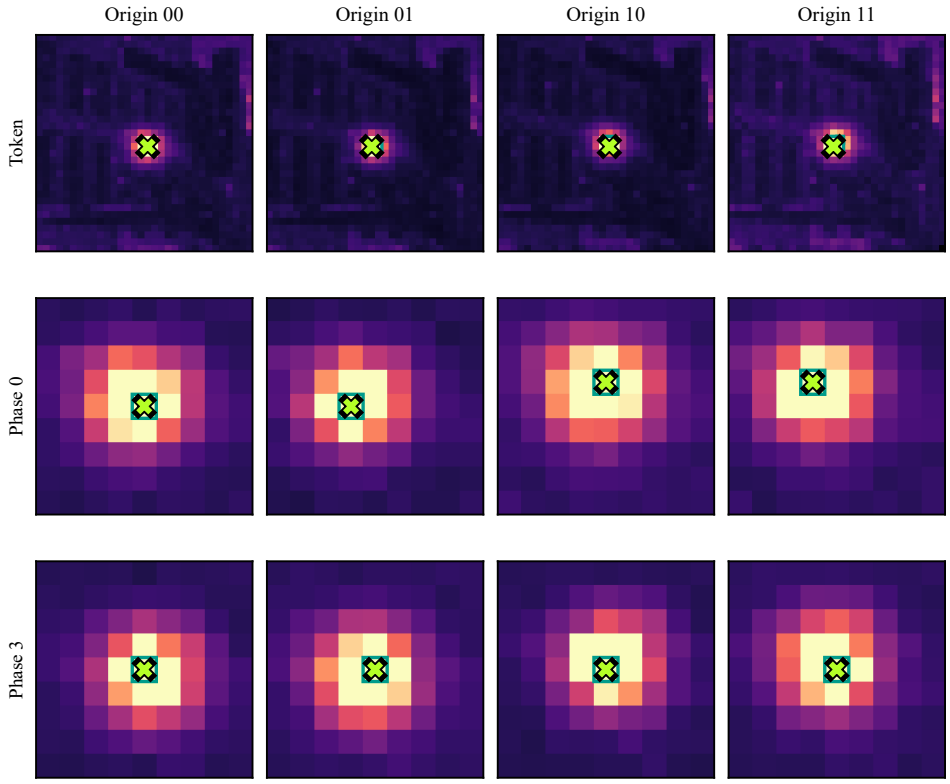


Figure 9: Predicted critical Haar windows and explicitly observed response peaks. Columns systematically permute the Haar origin while maintaining a fixed packed token intervention per row. Teal indicators denote predicted window indices; outlined lime crosses map measured high frequency peaks.

Population index consistency and functional head control. The predicted window test encompasses 36 deterministic dataset identities spanning distinct phases and origins. Figure 10(a) quantifies the response. Across all single phase conditions, $1124/1152 = 97.57\%$ of measured maxima localize perfectly within the pre specified window coordinates, firmly verifying the index alignment architecture.

Figure 10(b) tests functional control atop the distinct learned head output layer. Applying matched norm interventions upon token supports selected via high, random, or low gradient attribution yields mean local score changes of 1.045, 0.233, and 0.002 respectively, demonstrating exact structural regulatory control.

C.3 PTR: Inference Interventions and Explicit Cases

Correction magnitude and spatial response associations. Figure 11(a) scales the frozen PTR correction vector by coefficient $\alpha \in \{0, 0.25, 0.5, 0.75, 1, 1.25\}$, where $\alpha = 0$ represents a complete

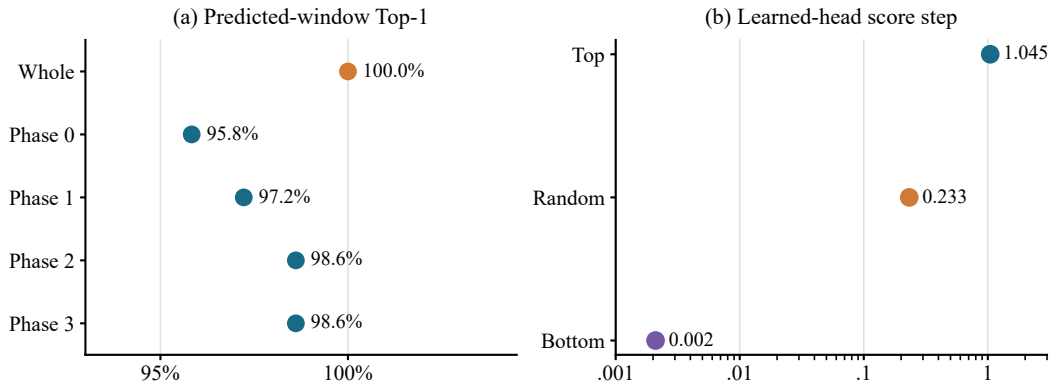


Figure 10: Level 1 architectural controls. (a) Pre specified critical Haar window Top 1 rates for global token and targeted single phase interventions. (b) Score response variance following matched norm structural interventions upon the token support tensor.

architectural bypass. The learned scaling factor $\alpha = 1$ maximizes validation fidelity, raising PSNR by 0.1463 dB relative to the bypass condition across 323 paired identities.

To isolate spatial behavior, the mechanism projects correction magnitude onto the RGB grid, partitioning pixels into magnitude deciles to measure absolute pixel MAE gain against the ground truth target. Figure 11(b,c) confirms that $298/323$ tested identities manifest distinct positive enrichment in regions receiving high correction magnitudes, explicitly linking the network's phase adjustment to objective local image recovery.

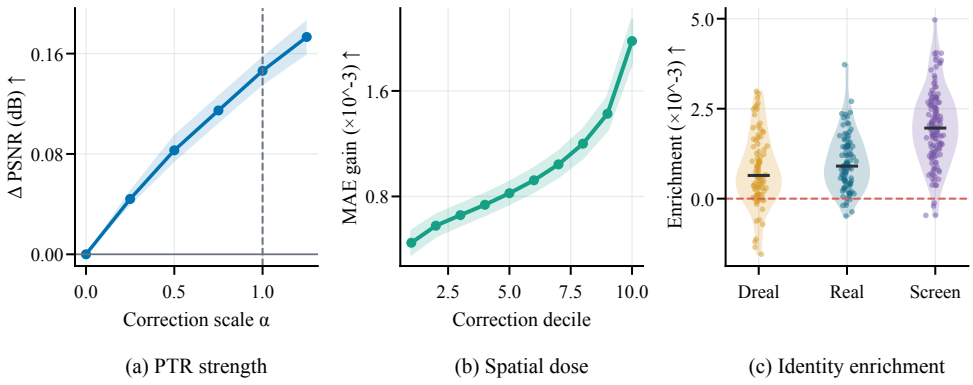


Figure 11: PTR correction strength and spatial associations across 323 validation identities. (a) Paired PSNR shift governed by correction scale α. (b) Positive pixel MAE gain distribution across increasing correction deciles. (c) Identity level statistical enrichment comparing explicit high correction versus low correction subregions.

Visualizing exact pathway responses. The following structural interventions isolate the effects of individual correction pathways on representative scenes. These selected cases satisfy strict improvement criteria ($\Delta\text{PSNR} > 0$ and $\Delta\text{PAI} < 0$ for Full versus Bypass) to explicitly visualize the targeted high frequency restoration mechanisms, complementing the aggregated population metrics. Each array isolates absolute output variance induced by eliminating DC, Contrast, or Boundary interaction paths, mapping precise physical network reactions.

D Detailed Experimental Setups

Datasets and Metrics. Training utilizes the DIV2K and LSDIR datasets. Real world degradation characteristics are applied via the Real-ESRGAN pipeline. Evaluation encompasses three paired

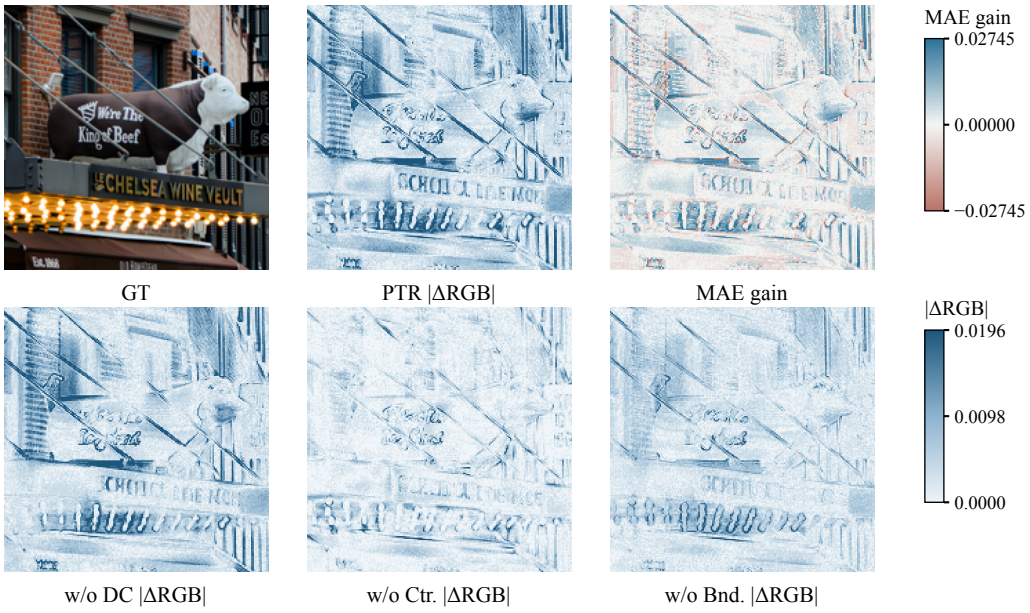


Figure 12: PTR spatial configuration 1: `ScreenSR_003`. A structured domain defined by robust signage geometries. Positive intensity in the pathway difference maps outlines exact architectural sensitivity.

benchmarks (RealSR, DrealSR, ScreenSR) and three unpaired environmental datasets (RealLR200, RealLQ250, RealPhoto60). Quantitative metrics execute through `pyiqa`. Full reference fidelity employs PSNR, SSIM, LPIPS, DISTS, and A-FINE. Reference free perceptual evaluation utilizes MUSIQ, MANIQA, CLIPIQA+, TOPIQ-NR, LIQE, Q-Align, and HyperIQA.

Compared Methods. The robust benchmark incorporates U-Net foundations (ResShift, SeeSR, SUPIR, SinSR, OSEDiff, HYPIR) and prominent DiT foundations (DiT4SR, TSD-SR, ODTSR, StrSR). These encompass iterative sampling trajectories alongside accelerated single step generative priors.

Basic Implementation Details. The architecture instantiates the FLUX.2 [klein] foundation, operating deterministically at a fixed timestep $t_\star = 500$ ($\sigma_\star \approx 0.5$). Optimization updates the VAE encoder, DiT velocity predictor, and PTR mechanism via LoRA (rank 64), preserving a completely frozen VAE decoder. The SPP references derive offline via the frozen initialization prior using perturbation $\sigma = 0.05$. Optimization executes across 10,000 global steps upon 2 NVIDIA A100 processors, specifying a batch size of 1 per device supplemented by 4 gradient accumulation stages. Objective scaling parameters enforce $\lambda_{\text{blur}} = 0.5$, $\lambda_{\text{perc}} = 3.0$, and $\lambda_{\text{adv}} = 0.3$.

Training Configurations. The processing pipeline extracts 512×512 spatial crops, evaluating gradients natively in bfloat16 mixed precision. Inference evaluates a static restoration prompt, avoiding independent low quality conditioning variants. The optimizer instantiates AdamW leveraging a constant learning rate of 5×10^{-5}, maintaining structural momentum parameters $(\beta_1, \beta_2) = (0.9, 0.999)$ and decay factors at 10^{-4}. An exponential moving average tracks generator weights with a momentum ratio of 0.999.

Loss Implementation Specifics. The perceptual objective $\mathcal{L}_{\text{perc}}$ combines representations from a frozen ConvNeXtV2-Huge backbone, applying uniform weight across the native spatial DISTS distance and the aggregated DISTS response derived from four distinct shifted Haar representations. The low frequency constraint integrates a timestep dependent spatial Gaussian blur. For a specified timestep t, the structural kernel spans $5 + 4\lfloor t/200 \rfloor$ with an associated standard deviation of $1 +$

$t/200$. Operating precisely at $t_\star = 500$, the configuration dictates a blur kernel size measuring 13 with a standard deviation of 3.5.

TPWD Multiple Shift Sampling. During the adversarial gradient update, exactly one structural Haar grid offset o determines spatial geometry per iteration. This shift coordinate enforces symmetry across real representations, generated distributions, and all competing network updates. Grid sampling maintains strict uniform probability:

$$\Pr(o) = \frac{1}{4} \quad \text{for all } o \in \{(0,0),(0,1),(1,0),(1,1)\}. \tag{14}$$

D.1 Numerical Quality Along the Frozen Base Flow

To isolate the inherent generative prior, we evaluate the frozen foundation model across 5000 ground truth (GT) images. Each image is encoded into the latent space, perturbed with Gaussian noise at scale σ, and restored via a single endpoint prediction. We report full reference fidelity (PSNR, LPIPS) alongside a normalized consensus effect derived from six reference free perceptual metrics (CLIPIQA, CLIPIQA+, MUSIQ, MANIQA, TOPIQ-NR, and LIQE). For each metric, the raw deviation from the GT score is scaled by the sample standard deviation; these scaled deviations are averaged to compute the final consensus effect. Table 8 reports the sample means and 95% bootstrap intervals.

Table 8: Restoration quality along the frozen base flow across varying noise scales σ. The consensus effect aggregates normalized deviations of six reference free perceptual metrics. Intervals represent 95% bootstrap bounds.

σ	NR consensus effect	95% interval	PSNR ↑	LPIPS ↓
6.18×10^{-6}	$+0.0794$	$[+0.0625, +0.0972]$	33.405	0.0166
1.35×10^{-4}	$+0.0788$	$[+0.0624, +0.0959]$	33.406	0.0166
1.04×10^{-3}	$+0.0783$	$[+0.0614, +0.0958]$	33.406	0.0166
9.67×10^{-3}	$+0.0790$	$[+0.0618, +0.0968]$	33.376	0.0167
0.1	$+0.0937$	$[+0.0741, +0.1129]$	31.823	0.0233
0.2	$+0.1044$	$[+0.0832, +0.1253]$	29.317	0.0416
0.4	$+0.0520$	$[+0.0097, +0.0926]$	25.226	0.1083
0.8	-2.1902	$[-2.4711, -1.9230]$	19.266	0.4903

At $\sigma = 0.1$, the absolute means of all six reference free metrics strictly exceed their respective GT baselines (e.g., MANIQA increases from 0.7064 to 0.7178, and CLIPIQA from 0.7089 to 0.7221). While the maximum perceptual consensus emerges at $\sigma = 0.2$, the corresponding LPIPS distance nearly doubles compared to $\sigma = 0.1$ (0.0416 versus 0.0233). Further increasing the noise scale to $\sigma = 0.4$ preserves a positive consensus effect but incurs substantial structural distortion. At $\sigma = 0.8$, the generative trajectory collapses, driving all reference free metric means below the GT baseline.

These measurements empirically establish a bounded, metric positive neighborhood for frozen model interventions. The trajectory confirms that modest noise perturbations induce a distinct perceptual enhancement phase, identifying the $\sigma \leq 0.1$ regime as a reliable target for constructing training references without sacrificing geometric fidelity.

E Methodological Details

E.1 PTR Implementation Details

Polyphase representation. Let ρ denote the invertible rearrangement that exposes implicit token phases from the packed velocity prediction P^0:

$$Q = \rho(P^0) \in \mathbb{R}^{h \times w \times 4 \times C}, \qquad Q_{i,j,q} = \left[\Pi^{-1}(P^0)\right]_{2i+d_y^q,\, 2j+d_x^q}, \tag{15}$$

where $q \in \{0,1,2,3\}$ indexes the four token phases with offsets (d_y^q, d_x^q) in row-major order. The rearrangement ρ retains separate packing cell and phase axes, whereas Π^{-1} places the identical

entries on the $2h \times 2w \times C$ latent grid. Each phase undergoes normalization and linear projection, subsequently incorporating a fixed positional encoding for token phases e_q:

$$H^{(0)}_{i,j,q} = W_{\text{in}}(\text{Norm}(Q_{i,j,q})) + e_q. \tag{16}$$

Phase Input and Fixed Positional Encoding. For the token phase component $Q_{i,j,q} \in \mathbb{R}^C$ retrieved from Eq. (15), PTR applies RMSNorm and a shared affine projection. For a feature array $u \in \mathbb{R}^{d_u}$,

$$\text{RMSNorm}_g(u) = g \odot \frac{u}{\sqrt{d_u^{-1} \sum_{r=1}^{d_u} u_r^2 + \varepsilon}}, \tag{17}$$

where g represents a learned channel gain. The phase input initialization evaluates as

$$H^{(0)}_{i,j,q} = W_{\text{in}}\, \text{RMSNorm}_{g_{\text{patch}}}(Q_{i,j,q}) + b_{\text{in}} + e_q. \tag{18}$$

The positional encoding e_q utilizes a deterministic functional formulation. For the row-major phases $(d_y^q, d_x^q) \in \{(0,0),(0,1),(1,0),(1,1)\}$, defining variables $a_q = d_y^q/2 - 1/4$ and $b_q = d_x^q/2 - 1/4$ yields:

$$e_q = \text{repeat}_{d/4}\,[a_q,\, b_q,\, a_q b_q,\, 1] \in \mathbb{R}^d. \tag{19}$$

Intra Cell Attention. Defining $U_{i,j,q} = \text{RMSNorm}_{g_{\text{intra}}}(H^{(0)}_{i,j,q})$, a shared affine map produces query, key, and value features:

$$[\widetilde{Q}_{i,j,q}, \widetilde{K}_{i,j,q}, \widetilde{V}_{i,j,q}] = W_{\text{qkv}} U_{i,j,q} + b_{\text{qkv}}, \tag{20}$$

which bifurcate into $m = 4$ heads of dimension $d_h = d/m = 16$. For head a, query phase q, and key phase r situated in the identical packing cell,

$$\begin{aligned} s^a_{q,r} &= \frac{\langle \widetilde{Q}^a_{i,j,q}, \widetilde{K}^a_{i,j,r} \rangle}{\sqrt{d_h}} + B^{\text{in}}_{a,\, d_y^r - d_y^q + 1,\, d_x^r - d_x^q + 1}, \\ \alpha^a_{q,r} &= \frac{\exp(s^a_{q,r})}{\sum_{r'=0}^{3} \exp(s^a_{q,r'})}. \end{aligned} \tag{21}$$

The intra cell feature update follows as:

$$H^{(1)}_{i,j,q} = H^{(0)}_{i,j,q} + W_o\, \text{Concat}_{a=1}^{m} \left(\sum_{r=0}^{3} \alpha^a_{q,r} \widetilde{V}^a_{i,j,r} \right) + b_o. \tag{22}$$

Cross Boundary Attention. Establishing $V_{i,j,q} = \text{RMSNorm}_{g_{\text{boundary}}}(H^{(1)}_{i,j,q})$, the mechanism maps phase (i,j,q) to the fine grid coordinate $(r,c) = (2i + d_y^q, 2j + d_x^q)$. The candidate neighborhood $\mathcal{N}_{\text{bd}}(r,c)$ incorporates directional relations $\tau(r,c;n) \in \{\text{self}, \text{up}, \text{down}, \text{left}, \text{right}\}$:

$$q^a_{r,c} = W^a_Q V_{r,c} + b^a_Q, \qquad k^a_n = W^a_K V_n + b^a_K, \qquad v^a_n = W^a_V V_n + b^a_V. \tag{23}$$

For every valid neighbor $n \in \mathcal{N}_{\text{bd}}(r,c)$,

$$\begin{aligned} s^a_{r,c,n} &= \frac{\langle q^a_{r,c}, k^a_n \rangle}{\sqrt{d_h}} + B^{\text{bd}}_{a,\tau(r,c;n)}, \\ \beta^a_{r,c,n} &= \frac{\exp(s^a_{r,c,n})}{\sum_{n' \in \mathcal{N}^{\text{valid}}_{\text{bd}}(r,c)} \exp(s^a_{r,c,n'})}. \end{aligned} \tag{24}$$

GEGLU Refinement. Following boundary interaction, a shared phase wise GEGLU refines the hidden state array. Let $Z_{i,j,q} = \text{RMSNorm}_{g_{\text{mlp}}}(H^{(2)}_{i,j,q})$. With hidden width $d_{\text{mlp}} = 2d$,

$$\begin{aligned} [A_{i,j,q}, G_{i,j,q}] &= W_1 Z_{i,j,q} + b_1, \\ \mathcal{M}(Z_{i,j,q}) &= W_2(A_{i,j,q} \odot \text{GELU}(G_{i,j,q})) + b_2, \\ H_{i,j,q} &= H^{(2)}_{i,j,q} + \mathcal{M}(Z_{i,j,q}). \end{aligned} \tag{25}$$

E.2 TPWD Implementation Details

For an input feature F and grid offset $o \in \{(0,0),(0,1),(1,0),(1,1)\}$, the implementation shifts the Haar grid by o, utilizing replicate padding to complete the border blocks before applying the signed 2×2 Haar transform. Let $\mathcal{C}_o(F) = \mathrm{Concat}(F_o^{\mathrm{LL}}, F_o^{\mathrm{LH}}, F_o^{\mathrm{HL}}, F_o^{\mathrm{HH}})$. The normalized representation evaluates as

$$\mathcal{H}_o(F) = \mathcal{U}\left(\frac{\mathcal{C}_o(F)}{\sqrt{\mathrm{Mean}(\mathcal{C}_o(F)^2) + 10^{-6}}}\right), \tag{26}$$

where $\mathcal{U}$ denotes nearest neighbor interpolation back to the original feature map resolution. At each distinct feature level $\ell \in \{1,2,3\}$, the discriminator heads D_ℓ^{sem}, D_ℓ^{LL}, and D_ℓ^{HF} apply stride 1 binomial BlurPool, a spectrally normalized 3×3 stride 2 convolution, LeakyReLU, and a spectrally normalized 1×1 convolution.

Let $h_b \in \mathbb{R}^4$ represent the row major vectorization of K^b in Eq. (10). The four vectors construct an orthonormal basis, maintaining $h_{\mathrm{LL}} = \frac{1}{2}\mathbf{1}$. Consequently, the PTR projectors precisely mirror the Haar decomposition structure:

$$\mathbf{P}_{\mathrm{dc}} = h_{\mathrm{LL}} h_{\mathrm{LL}}^\top, \qquad \mathbf{P}_{\mathrm{ctr}} = \sum_{b\in\{\mathrm{LH,HL,HH}\}} h_b h_b^\top. \tag{27}$$

E.3 Ideal Flow Interpretation of Self Prior Projection

Within the flow matching framework, considering the linear interpolation $Z_\sigma = (1-\sigma)Z + \sigma\epsilon$, where $Z \sim p_{\mathrm{base}}^z$ represents the latent distribution of the pretrained foundation model and $\epsilon \sim \mathcal{N}(0,I)$. The optimal marginal velocity prediction $v^\star(u,\sigma) = \mathbb{E}[\epsilon - Z \mid Z_\sigma = u]$ satisfies

$$\begin{aligned} u - \sigma v^\star(u,\sigma) &= \mathbb{E}[Z_\sigma - \sigma(\epsilon - Z) \mid Z_\sigma = u] \\ &= \mathbb{E}[Z \mid Z_\sigma = u]. \end{aligned} \tag{28}$$

Subtracting the source latent results in $z^+ - z = \sigma\left[\epsilon - z - v_0(z_\sigma,\sigma)\right]$. For a fixed source latent and noise realization, a bounded local field response guarantees $\|z^+ - z\| = \mathcal{O}(\sigma)$. Maintaining a minor perturbation level preserves the proximity of the projected latent to the encoded source while enabling the frozen prior to dictate the localized displacement vector.

E.4 Shared and Relative Decomposition Scales

Let $h_b \in \mathbb{R}^4$ formulate the row major vectorization of K^b. The four variables formulate an orthonormal basis, assuring $h_{\mathrm{LL}} = \frac{1}{2}\mathbf{1}$. PTR coordinates these exact shared and relative subspaces for latent velocity correction; TPWD applies the identical decomposition upon local image feature grids. This symmetry mathematically links the structural correction principles across both prediction and penalty domains.

The nominal scale ratio between a 2×2 feature block and a 2×2 packing cell equals $s_\ell / r_{\mathrm{VAE}}$. A Haar grid offset o directly triggers a nominal image displacement $s_\ell o$ and a corresponding latent displacement $(s_\ell / r_{\mathrm{VAE}})o$. Utilizing a decoder factor $r_{\mathrm{VAE}} = 8$, feature strides of 4, 8, and 16 instantiate the following sampling grid correspondence:

Feature stride	Haar block span	Relative linear scale	Latent shift per feature pixel
4	8×8	$1/2$	$1/2$
8	16×16	1	1
16	32×32	2	2

At matched scale and aligned origins, shifted Haar blocks explicitly traverse packing cell boundaries, ensuring sub token and cross token geometry evaluations.

F LIMITATIONS

While PhaDiSR effectively mitigates token phase dependence and establishes prominent advantage in reference free perceptual benchmarks, this dual domain correction paradigm introduces specific considerations. First, structural phase correction explicitly optimizes high frequency visual fidelity, a target which inherently diverges from pure pixel level smoothing objectives (e.g., PSNR). Specifically, eliminating constraints such as SPP demonstrably biases the generative output toward pixel level smoothness, severely suppressing essential perceptual textures. Second, while raw structural metrics naturally correlate with base model capacity, the extensive cross architecture benchmark robustly verifies that phase aware correction operates consistently as a foundational multiplier across varied DiT parameters. Future research trajectories could mathematically explore adaptive tokenization mechanisms that natively relax phase rigidity immediately upon latent extraction, prior to deep network evaluations.