DIVIDE AND LIGHTEN: LIGHTWEIGHT PIXEL DIFFUSION TRANSFORMERS FOR ONE-STEP IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

One-step diffusion substantially reduces the denoising cost of super-resolution, yet most existing methods still rely on latent diffusion architectures and thus remain burdened by expensive VAE encoding and decoding. We therefore propose LiPDiT-SR, a Lightweight Pixel Diffusion Transformer that eliminates the VAE and performs one-step super-resolution directly in pixel space. LiPDiT-SR first builds a full VAE-free one-step PixelDiT model for super-resolution, where the patch-level and pixel-level pathways primarily handle semantic reasoning and texture reconstruction, respectively. We then derive a compact model by pruning the computationally dominant patch-level pathway while preserving the pixel-level pathway. Modulation Deviation Block Selection (MDBS) selects the patch blocks whose removal least disrupts the modulation signals delivered to the pixel-level pathway, while a Semantic Modulation Alignment (SMA) loss aligns these signals with those of the preceding reference model at each pruning stage. Experiments on multiple synthetic and real-world benchmarks demonstrate that LiPDiT-SR achieves visual quality comparable to state-of-the-art one-step diffusion SR methods while delivering a 1.6$\times$ inference speedup for $\times 4$ super-resolution.

1 INTRODUCTION

Diffusion models (Li et al., 2022; Saharia et al., 2023) have demonstrated strong perceptual recovery capabilities for image super-resolution (SR), as their powerful natural image generation priors help synthesize realistic textures. However, diffusion models typically rely on multi-step iterative sampling, whose substantial computational cost limits their practical application (Wang et al., 2024b; Wu et al., 2024a).

Recent methods (Wang et al., 2024b; Wu et al., 2024a; You et al., 2025b; Dong et al., 2025) compress multi-step diffusion SR into a one-step mapping, but two efficiency bottlenecks remain. First, most inherit latent diffusion architectures (Rombach et al., 2022), whose pretrained VAE encodes the input and decodes the restored latent. These costs remain unchanged when denoising is reduced to one step and therefore occupy a substantial fraction of inference time (Wang et al., 2026) (Fig. 1). Second, generic compression treats layers and modules as interchangeable pruning units, overlooking their different roles in global structure and high-frequency detail synthesis. Multi-step diffusion progressively recovers coarse structure and fine textures, whereas a single forward pass must perform both simultaneously. Function-agnostic pruning may therefore remove computation essential to perceptual quality. Since the LR observation already supplies substantial semantic and structural information, SR offers a more targeted opportunity: compress the pathway responsible for global modeling while preserving the components critical to fine-detail synthesis.

PixelDiT (Yu et al., 2026) provides a suitable backbone: it operates directly in pixel space without VAE overhead and separates a patch-level pathway for global semantics from a pixel-level pathway for fine textures under pixel-wise AdaLN conditioning. We adapt a pretrained PixelDiT to one-step SR to obtain the full PixelDiT-SR reference. Profiling reveals that patch blocks account for 77.2% of inference time, compared with only 19.5% for the pixel-level pathway (Table 3). We therefore compress the patch blocks while preserving the pixel-level pathway and its pixel-wise AdaLN projections.

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.

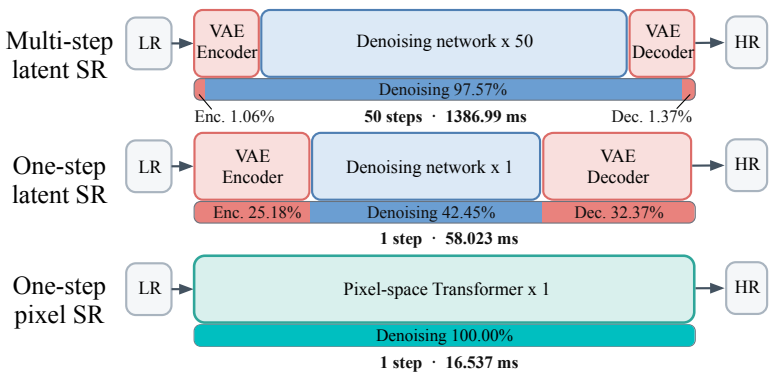


Figure 1: Stage-wise inference time breakdown of latent-space and pixel-space SR pipelines. The latent-space pipelines correspond to TSD-SR (Dong et al., 2025), and the pixel-space pipeline corresponds to our LiPDiT-SR.

We propose LiPDiT-SR, a lightweight one-step pixel diffusion Transformer for SR. It compresses the full reference's patch-level pathway while keeping the pixel-level pathway structurally complete and trainable. Modulation Deviation Block Selection (MDBS) ranks blocks by the change in pixel-level conditioning when each block is bypassed. Semantic Modulation Alignment (SMA) then aligns this conditioning with the preceding reference during progressive pruning. Both operate only during training and add no inference computation. Across synthetic and real-world benchmarks, LiPDiT-SR retains competitive quality while delivering a **1.6**$\times$ inference speedup for $\times 4$ SR.

Our main contributions are summarized as follows:

- We propose LiPDiT-SR, a one-step pixel diffusion Transformer for single-image super-resolution. Starting from a full PixelDiT-SR reference, it compresses the cost-dominant patch-level pathway while retaining the complete pixel-level renderer.
- We introduce MDBS and SMA to enable function-aware compression of the patch-level pathway. MDBS identifies and retains the patch blocks that most strongly influence the conditioning delivered to the pixel-level pathway, whereas SMA aligns the conditioning signals of the reference and current models at successive pruning stages.
- Experiments on multiple benchmarks show competitive visual quality and a **1.6**$\times$ inference speedup over SOTA one-step diffusion SR methods under $\times 4$ scaling.

2 Related Work

2.1 One-Step Diffusion for Image Super-Resolution

Diffusion models have recently demonstrated remarkable perceptual recovery capabilities in image super-resolution. StableSR (Wang et al., 2024a), DiffBIR (Lin et al., 2024), PASD (Yang et al., 2024), SeeSR (Wu et al., 2024b), and SUPIR (Yu et al., 2024) adapt pretrained text-to-image diffusion models to SR, but these methods typically require multiple calls to the denoising network, resulting in considerable computational cost. Most of them first map the LR image into the latent space of a pretrained VAE and then condition the denoising network through cross-attention or feature adapters.

To alleviate this overhead, recent studies compress multi-step diffusion SR into a one-step mapping. SinSR (Wang et al., 2024b) distills a pretrained ResShift (Yue et al., 2023) teacher into a deterministic one-step mapping, while OSEDiff (Wu et al., 2024a) starts diffusion directly from the LQ latent with variational score distillation. CTMSR (You et al., 2025b) learns the one-step process via consistency trajectory matching, InvSR (Yue et al., 2025) enables arbitrary-step sampling from an intermediate noisy state, and TSD-SR (Dong et al., 2025) and PiSA-SR (Sun et al., 2025) further improve texture quality via target-score modeling and pixel-semantic alignment. Despite single-step sampling, most methods retain a latent-space pipeline in which VAE encoding and decoding dom-

inate end-to-end latency. They also differ in how the LR condition is injected: SinSR and CTMSR train a deterministic LR-to-HR mapping, whereas OSEDiff and InvSR keep a score-based objective evaluated only once.

2.2 Pixel-Space Diffusion Transformers

Pixel-space diffusion avoids autoencoder overhead but must efficiently model high-dimensional images. HDiT (Crowson et al., 2024) uses a hierarchical architecture, while DiP (Chen et al., 2026) and DeCo (Ma et al., 2026b) decouple global structure or low-frequency semantics from local detail. PixelDiT (Yu et al., 2026) separates global modeling in a patch-level pathway from per-pixel rendering in a pixel-level pathway. This division allows different computational budgets for the two roles. It is particularly suitable for SR, whose LR condition already supplies low-frequency content, allowing compression of semantic processing while preserving detail synthesis.

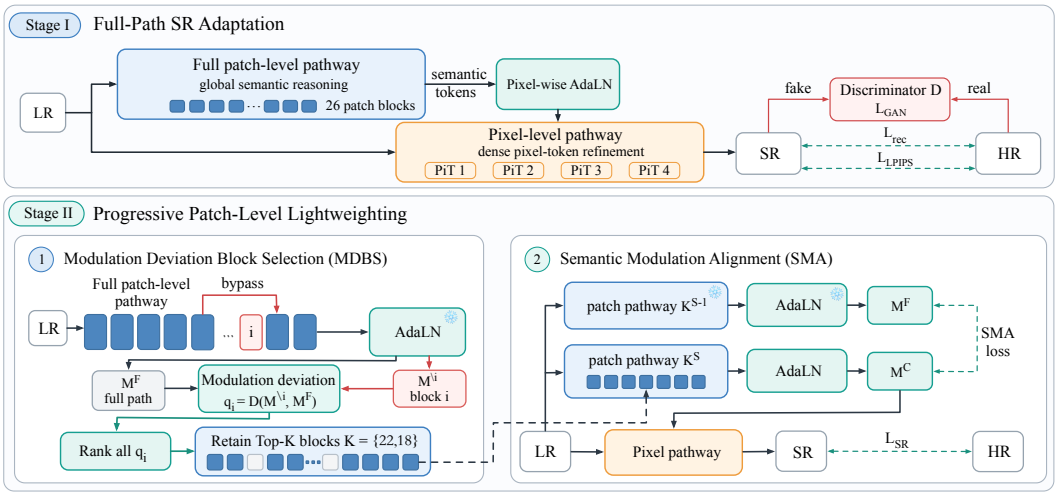


Figure 2: Overall framework of LiPDiT-SR. Stage I adapts pretrained multi-step PixelDiT into a full one-step SR reference. Stage II uses MDBS to progressively retain patch blocks and SMA to transfer modulation from the preceding K_{s-1} model to the current K_s model.

2.3 Diffusion Model Lightweighting for Image Super-Resolution

Beyond reducing sampling steps, general diffusion models can also be made cheaper per call: DiT-FastAttn (Yuan et al., 2024) reduces attention cost with local windows and cross-timestep computation sharing, DiffCR (You et al., 2025a) lets unimportant tokens bypass DiT layers with learned adaptive ratios, and PPCL (Ma et al., 2026a) prunes redundant contiguous layer intervals in DiTs with teacher–student distillation. Their criteria therefore do not directly indicate which parts of the network are critical in a restoration setting, where a strong input condition already carries much of that information.

For diffusion-based SR, AdcSR (Chen et al., 2025) removes or prunes modules of OSEDiff and compensates the pruned VAE decoder with adversarial distillation, while TinySR (Dong et al., 2026) performs depth pruning on a one-step diffusion Transformer and slims down its VAE with channel pruning and lightweight separable convolutions. However, these approaches define lightweighting granularity by generic structural units such as layers, modules, or tokens, without distinguishing the roles of different pathways in semantic reasoning and texture generation. Unlike these strategies, we exploit PixelDiT's dual-level architecture to prune only the cost-dominant patch-level pathway, retain the pixel renderer, choose blocks by modulation deviation, and align their conditioning during compact-model training. A restoration-oriented criterion should instead ask which components supply structure and which ones render detail. In particular, AdcSR removes the VAE encoder and text-side modules entirely, whereas TinySR keeps a shallow backbone that still synthesizes strong textures.

3 PROPOSED METHOD

3.1 OVERVIEW

PixelDiT (Yu et al., 2026) is a VAE-free pixel diffusion Transformer with a dual-level architecture. Given an LR input $\mathbf{y}$ and the time condition t, the full PixelDiT-SR model (Fig. 2) first obtains the semantic tokens $\mathbf{S}^F$ through the patch-level pathway $\mathcal{S}_F$ of L_s patch blocks.

The pixel-level pathway $\mathcal{R}_{\text{pix}}$, composed of L_p PiT blocks, then takes the bicubic-upscaled LR image $\mathbf{y}_\uparrow$ as input and generates the fine details. For the j-th PiT block, its pixel-wise AdaLN projection $\mathcal{A}_j$ converts the semantic tokens into the scale, shift, and residual-gating signals used by the attention and MLP branches, whose collection we denote by $\mathbf{M}_j$. We use the superscripts F and C to distinguish the full and lightweight models, respectively:

$$\mathbf{S}^F = \mathcal{S}_F(\mathbf{y}, t). \tag{1}$$

$$\mathbf{M}_j^F = \mathcal{A}_j(\mathbf{S}^F, t) = \{\boldsymbol{\gamma}_j^a, \boldsymbol{\beta}_j^a, \mathbf{g}_j^a, \boldsymbol{\gamma}_j^m, \boldsymbol{\beta}_j^m, \mathbf{g}_j^m\}, \tag{2}$$

$$\mathrm{AdaLN}(\mathbf{h}; \boldsymbol{\gamma}, \boldsymbol{\beta}) = (1 + \boldsymbol{\gamma}) \odot \mathrm{LN}(\mathbf{h}) + \boldsymbol{\beta}, \tag{3}$$

$$\hat{\mathbf{x}}^F = \mathcal{R}_{\text{pix}}\left(\mathbf{y}_\uparrow, \{\mathbf{M}_j^F\}_{j=1}^{L_p}\right), \tag{4}$$

where $\mathbf{h}$ denotes a pixel feature and LN denotes layer normalization, while the residual gate $\mathbf{g}$ scales the corresponding attention or MLP branch before its residual update.

LiPDiT-SR removes blocks only from the cost-dominant patch-level pathway; the pixel-level pathway remains structurally complete and trainable. Two challenges arise: block importance is not determined by depth, and removal changes downstream scale, shift, and gate signals. MDBS therefore selects blocks by their influence on this conditioning, while SMA preserves the retained pathway's compatibility with the pixel-level pathway.

3.2 MODULATION DEVIATION BLOCK SELECTION (MDBS)

MDBS considers a full patch-level pathway composed of sequential blocks $\{\mathcal{B}_i\}_{i=1}^{L_s}$. To estimate how the i-th block affects the downstream conditioning of the pixel-level pathway, we temporarily bypass that block on a fixed calibration set while leaving all other components unchanged. This produces perturbed semantic tokens $\mathbf{S}^{\setminus i}$ and the corresponding PiT modulation $\mathbf{M}_j^{\setminus i}$:

$$\mathbf{S}^{\setminus i} = \mathcal{S}_F^{\setminus i}(\mathbf{y}, t), \qquad \mathbf{M}_j^{\setminus i} = \mathcal{A}_j(\mathbf{S}^{\setminus i}, t). \tag{5}$$

Because scale, shift, and gate may have different numerical ranges, directly summing their absolute errors can cause one component to dominate the score. We therefore measure each modulation type using an ℓ_1 distance normalized by the magnitude of the corresponding full-path signal, with a small constant ϵ for numerical stability, and then average over PiT blocks, modulation types, and calibration samples:

$$d(\mathbf{M}_{j,u}^{\setminus i}, \mathbf{M}_{j,u}^F) = \frac{\|\mathbf{M}_{j,u}^{\setminus i} - \mathbf{M}_{j,u}^F\|_1}{\|\mathbf{M}_{j,u}^F\|_1 + \epsilon}, \tag{6}$$

$$q_i = \mathbb{E}_{\mathbf{y}}\left[\frac{1}{L_p|\mathcal{U}|}\sum_{j=1}^{L_p}\sum_{u\in\mathcal{U}} d(\mathbf{M}_{j,u}^{\setminus i}, \mathbf{M}_{j,u}^F)\right], \tag{7}$$

where $\mathcal{U} = \{\gamma^a, \beta^a, g^a, \gamma^m, \beta^m, g^m\}$ and q_i is the modulation deviation score used by MDBS. A larger q_i indicates that bypassing block i causes a larger change in the conditioning signals actually consumed by the pixel-level pathway and that the block is therefore more important to it.

Given a target budget of K blocks, MDBS retains the blocks with the highest modulation deviation scores, where TopK returns the indices of the K largest scores, and preserves their original relative

order to construct the lightweight patch-level pathway $\mathcal{S}_C$. Selection is performed once offline before lightweight-model training. It requires neither a learned search network nor dynamic routing or scoring during inference. The same ordering also drives the progressive pruning schedule in Stage II below: every budget retains a prefix of the ordering, so that the retained block sets are nested across steps:

$$\mathcal{I}_K = \mathrm{TopK}(\{q_i\}_{i=1}^{L_s}), \qquad \mathbf{S}^C = \mathcal{S}_C(\mathbf{y}, t; \mathcal{I}_K). \tag{8}$$

3.3 Semantic Modulation Alignment (SMA)

Selecting blocks according to q_i alone does not guarantee that the current lightweight pathway will remain compatible with the model from which it is initialized. The pixel-level pathway does not consume the semantic tokens; it consumes the modulation signals produced from them by AdaLN. SMA therefore preserves the conditioning passed from the patch-level to the pixel-level pathway instead of forcing the current pathway to reproduce every intermediate representation.

At progressive stage s, the current model produces $\mathbf{M}_j^{K_s}$, while the frozen preceding model provides $\mathbf{M}_j^{K_{s-1}}$ without gradient computation. SMA applies the normalized distance used by MDBS to every scale, shift, and gate signal across all PiT blocks. Stop-gradient keeps the preceding model detached, yielding the following alignment objective:

$$\mathcal{L}_{\mathrm{SMA}}^{(s)} = \frac{1}{L_p|\mathcal{U}|} \sum_{j=1}^{L_p} \sum_{u \in \mathcal{U}} \frac{\|\mathbf{M}_{j,u}^{K_s} - \mathrm{sg}(\mathbf{M}_{j,u}^{K_{s-1}})\|_1}{\|\,\mathrm{sg}(\mathbf{M}_{j,u}^{K_{s-1}})\|_1 + \epsilon}. \tag{9}$$

For $s = 1$, the full $K_0 = 26$ model is the reference; subsequent stages use the preceding compact model. Thus, Eq. (9) aligns each K_s model with its immediately preceding K_{s-1} model, preserving the conditioning consumed by the pixel-level pathway.

This interface-level target is important because semantic representations are not unique: different patch features can produce similar downstream behavior after the learned AdaLN projections. Direct feature matching therefore constrains information that may not affect pixel rendering, whereas image-level supervision alone does not control the shifted conditioning distribution. SMA allows the current model to reorganize its internal features while preserving the rendering instructions of the preceding model.

3.4 Overall Training Procedure

Stage I: Full-Path SR Adaptation. We first adapt the complete PixelDiT architecture to SR, allowing the pixel-level pathway to learn texture and local-detail generation for real-world SR and producing the full functional reference used by the subsequent lightweighting stage. We optimize this model with the joint SR objective below:

$$\mathcal{L}_{\mathrm{SR}} = \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{per}}\mathcal{L}_{\mathrm{LPIPS}} + \lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{GAN}}. \tag{10}$$

Stage II: Progressive Patch-Level Lightweighting. Instead of pruning directly to the target budget, we set the initial patch-level depth to $K_0 = 26$ and progressively reduce it according to $\{K_s\}_{s=1}^{S} = \{22, 18\}$. At stage s, we retain the top K_s blocks under the fixed MDBS ranking to construct the current model. The current model is initialized from the preceding K_{s-1} block model. We freeze the preceding model as the SMA reference. All components of the current K_s model are trainable. We optimize the current model with the image-level SR objective and an SMA term that aligns its conditioning with that of the reference:

$$\mathcal{L}_{\mathrm{stage2}}^{(s)} = \mathcal{L}_{\mathrm{SR}}(\hat{\mathbf{x}}^{K_s}, \mathbf{x}) + \lambda_m \mathcal{L}_{\mathrm{SMA}}^{(s)}, \tag{11}$$

where $\hat{\mathbf{x}}^{K_s}$ is the output of the current model and $\mathbf{x}$ is the HR target. The SR objective backpropagates through all components of the current model. The SMA term follows Eq. (9): its gradients pass

Table 1: Quantitative image quality results of different methods on ImageNet-Test, RealSR, and RealSet65. The best and second-best results are marked in bold and underlined. All results of previous methods are evaluated using the released codes and pre-trained weights.

Metric	SinSR	OSEDiff	InvSR	PiSA-SR	TSD-SR	CTMSR	AdcSR	Full PixelDiT-SR	Ours
ImageNet-Test									
PSNR↑	<u>27.502</u>	24.878	23.819	25.338	25.045	**27.539**	25.264	24.940	24.892
SSIM↑	<u>0.7522</u>	0.7032	0.6809	0.7005	0.6916	**0.7556**	0.7035	0.7435	0.7481
LPIPS↓	<u>0.2977</u>	0.3210	0.3311	0.3051	0.3177	**0.2854**	0.3302	0.3421	0.3392
CLIPIQA↑	0.6571	0.6811	0.7059	0.7142	<u>0.7516</u>	0.6999	0.7099	**0.7602**	0.7420
MUSIQ↑	68.577	72.027	72.934	73.546	**74.473**	71.323	72.735	72.066	71.818
MANIQA↑	0.4488	0.4678	0.4735	0.5136	**0.5551**	0.4925	0.5085	<u>0.5393</u>	0.4939
NIQE↓	5.040	4.429	4.480	<u>4.117</u>	**3.981**	4.188	4.321	4.548	4.554
RealSR									
PSNR↑	**26.645**	25.155	24.431	25.547	24.759	<u>26.398</u>	25.596	25.543	25.351
SSIM↑	0.7575	0.7363	0.7221	0.7442	0.7207	**0.7735**	0.7325	<u>0.7720</u>	0.7681
LPIPS↓	0.4215	0.3765	0.3890	**0.3644**	0.3939	0.3755	0.3867	<u>0.3714</u>	0.3792
CLIPIQA↑	0.6371	0.6771	0.6957	0.6862	<u>0.7290</u>	0.6116	0.7015	**0.7296**	0.6881
MUSIQ↑	59.855	68.061	67.750	68.691	**69.926**	63.303	67.595	66.816	66.167
MANIQA↑	0.3794	0.4670	0.4487	0.4830	<u>0.4977</u>	0.3934	**0.5068**	0.4428	0.4077
NIQE↓	5.025	4.257	<u>4.168</u>	4.173	**3.852**	4.659	4.306	5.039	4.826
RealSet65									
CLIPIQA↑	0.6628	0.6921	0.7097	0.7252	**0.7534**	0.6647	0.7055	<u>0.7352</u>	0.7053
MUSIQ↑	61.437	69.691	68.976	71.099	**71.642**	65.342	69.238	67.046	65.735
MANIQA↑	0.3911	0.4501	0.4632	**0.4937**	<u>0.4888</u>	0.4121	0.4654	0.4482	0.4054
NIQE↓	5.271	4.304	4.077	**4.030**	<u>4.064</u>	4.510	4.332	4.554	4.412

only through the current model's modulation-producing path, while the K_{s-1} reference remains detached. Progressive reduction limits the conditioning gap to be bridged during each joint-adaptation stage. The final stage yields LiPDiT-SR with $K_S = 18$ patch blocks.

The deployed model consists only of $\mathcal{S}_C$, $\{\mathcal{A}_j\}_{j=1}^{L_p}$, and $\mathcal{R}_{\text{pix}}$. The stage-wise reference models, MDBS scoring, and SMA loss do not participate in inference. Consequently, the runtime reduction of LiPDiT-SR comes directly from removing patch blocks rather than from an auxiliary approximation or dynamic execution mechanism.

4 Experiments

4.1 Experimental Setup

Datasets. We use 1.2 million images from the ImageNet training set (Deng et al., 2009) for training. Under the $\times 4$ SR setting, we randomly crop 512×512 HR patches and synthesize the corresponding 128×128 LR inputs using the degradation pipeline of Real-ESRGAN (Wang et al., 2021). For evaluation, we adopt ImageNet-Test (Deng et al., 2009), which contains 3,000 paired 512×512 images, as the synthetic degradation benchmark, whose LR inputs are generated with the same Real-ESRGAN degradation protocol used by existing one-step SR methods. We further evaluate the model on two real-world datasets, RealSR (Cai et al., 2019) and RealSet65 (Yue et al., 2023).

Implementation Details. We initialize the full model from the official PixelDiT-C2I weights (Yu et al., 2026) and remove its class conditioning. The full model contains $L_s = 26$ patch blocks and $L_p = 4$ PiT blocks. During training, Stage I unfreezes all parameters for SR adaptation and optimizes them for 100K iterations with an L1 reconstruction loss of $\lambda_{\text{rec}} = 1.0$ and an LPIPS perceptual loss of $\lambda_{\text{per}} = 0.5$; the hinge GAN weight λ_{adv} is linearly increased from 0 to 0.1 between 20K and 40K iterations and kept fixed afterwards. In Stage II, MDBS is first applied on a fixed calibration set to rank the patch blocks. Starting from $K_0 = 26$, the patch-level pathway is progressively pruned according to $\{K_s\}_{s=1}^{S} = \{22, 18\}$. At each stage s, the preceding K_{s-1} block model is kept fixed as the reference, while the K_s block model is initialized from it and optimized

with $\mathcal{L}_{\text{SR}} + \lambda_m \mathcal{L}_{\text{SMA}}^{(s)}$. Both stages are trained on a single NVIDIA RTX PRO 6000 GPU with a batch size of 16. In each stage, the generator and the discriminator are optimized with AdamW using a learning rate of 1×10^{-5}, betas of $(0.9, 0.99)$. After 2,500 warm-up steps, the learning rate is halved at 50K and 90K iterations.

Table 2: Inference efficiency on ImageNet-Test under unified $\times 4$ evaluation. Runtime excludes model loading and metric computation; memory is the average per-image peak reserved GPU memory. All released methods use the same timing protocol and hardware.

Method	ImageNet-Test			
	ms/image↓	Memory (GiB)↓	FLOPs (T)↓	Params (M)↓
SinSR	65.10	1.73	5.294	173.92
OSEDiff	70.94	3.97	4.832	1765.30
InvSR	60.40	3.16	5.809	1323.79
PiSA-SR	46.48	4.66	4.481	1289.95
TSD-SR	58.02	4.92	6.161	2207.33
CTMSR	39.05	**1.30**	1.226	**171.52**
AdcSR	26.29	2.19	1.099	456.07
LiPDiT-SR	**16.54**	2.51	**1.053**	605.04

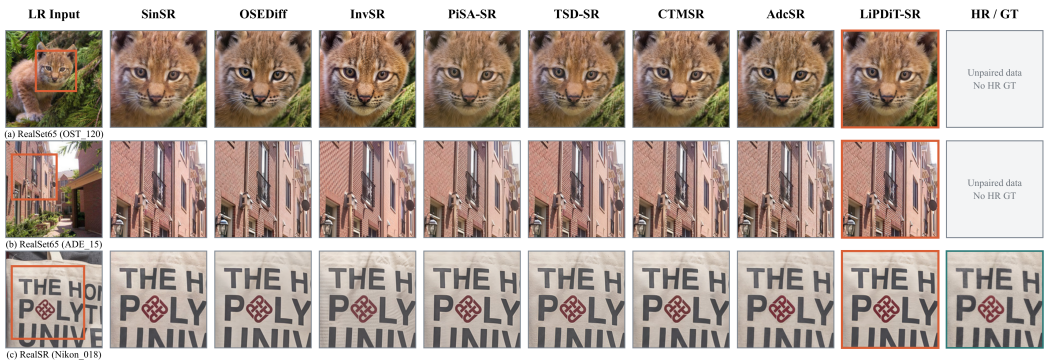


Figure 3: Qualitative comparison on synthetic and real-world degraded images.

Evaluation Metrics. We adopt fidelity-oriented and perception-oriented metrics to evaluate the results. PSNR and SSIM (Wang et al., 2004) are computed on the luminance channel. PSNR, SSIM, and LPIPS (Zhang et al., 2018) are full-reference metrics, whereas CLIPIQA (Wang et al., 2023), MUSIQ (Ke et al., 2021), MANIQA (Yang et al., 2022), and NIQE (Mittal et al., 2013) are no-

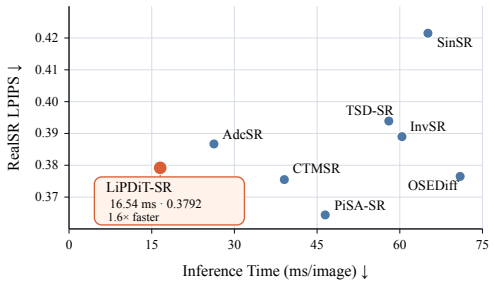


Figure 4: Quality–inference-time trade-off between RealSR LPIPS and batch-1 inference time measured on ImageNet-Test under the unified hardware and timing protocol. Lower-left is better.

Table 3: Module-level inference time breakdown of the full PixelDiT-SR under the 512×512 setting. Patch blocks form the patch-level pathway.

Module	Time (ms)	Share (%)
Patch blocks	17.50	77.2
Pixel-level pathway	4.43	19.5
Remaining components	0.74	3.3
Total	22.67	100

reference metrics. HR reference is not available for RealSet65, only the no-reference metrics are reported.

Runtime Profiling. Table 3 verifies the strong computational asymmetry that motivates our design. Patch blocks consume 17.50 of the full model's 22.67 ms latency, nearly four times the cost of the complete pixel-level pathway. Consequently, uniformly reducing both pathways would sacrifice pixel-rendering capacity while leaving the dominant semantic cost insufficiently addressed. Our selective strategy instead spends the limited compression budget where the measured latency is concentrated and preserves the substantially cheaper pathway that directly synthesizes local details.

4.2 Comparison with State-of-the-Art Methods

Our goal is not to optimize any single quality metric, but to reduce inference time while maintaining visual quality comparable to strong one-step SR methods.

Quantitative Visual Quality. As shown in Table 1, LiPDiT-SR delivers competitive visual quality across all three benchmarks. On ImageNet-Test, its fidelity and perceptual metrics remain close to those of the full PixelDiT-SR model, which achieves the best CLIPIQA on ImageNet-Test and RealSR (0.7602 and 0.7296), demonstrating the strong image prior of the pixel-space generator. Despite removing 8 of the 26 patch blocks, LiPDiT-SR exhibits no metric collapse and even improves ImageNet-Test LPIPS from 0.3421 to 0.3392. On RealSR, the corresponding changes in PSNR and SSIM are limited to 0.192 dB and 0.0039, respectively. Its competitive no-reference scores on the unpaired RealSet65 further indicate that the lightweight model generalizes beyond the synthetic training degradation. Together, MDBS and SMA preserve the pixel pathway's detail-generation capacity.

Inference Efficiency. We evaluate batch-1 latency, peak GPU memory, FLOPs, and parameter count on ImageNet-Test under unified $\times 4$ restoration on one NVIDIA RTX PRO 6000 GPU. Latency excludes model loading and metric computation; memory is the per-image average of peak reserved GPU memory. LiPDiT-SR uses 16.54 ms/image, 1.053 T FLOPs, and 2.51 GiB (Table 2). Relative to the fastest baseline, it reduces latency by 37.1% (26.29 to 16.54 ms; **1.6**$\times$ speedup) and improves RealSR LPIPS from 0.3867 to 0.3792. CTMSR is more memory-efficient and achieves LPIPS 0.3755 at 2.36$\times$ the latency; PiSA-SR reaches 0.3644 at 2.81$\times$. CTMSR's 171.52M parameters are also fewer than our 605.04M, so model size alone does not explain the speedup. Instead, LiPDiT-SR eliminates VAE execution, which exceeds half the cost of the one-step latent pipeline (Fig. 1), and prunes the pathway responsible for 77.2% of full-model latency (Table 3). It has the lowest FLOPs among the evaluated models, with the quality–time trade-off shown in Fig. 4.

Qualitative Comparison. To demonstrate the visual performance of our method, we compare LiPDiT-SR with other state-of-the-art one-step methods in Fig. 3. Its results remain visually close to the full PixelDiT-SR model. In structure-rich regions such as text edges and repetitive building textures, LiPDiT-SR reconstructs sharp boundaries and regular patterns, whereas several latent-space one-step methods oversmooth these structures or synthesize textures misaligned with the LR input. The animal example also shows coherent facial contours and directional fur rather than locally inconsistent high-frequency patterns. These observations confirm that preserving the pixel-level pathway and its conditioning interface retains the full model's detail-rendering ability, while reducing the patch-level pathway primarily removes redundant semantic computation.

4.3 Ablation Studies

Table 4: Ablation of MDBS and SMA on RealSR. All lightweight variants retain the same number of patch blocks and the complete pixel-level pathway.

Variant	Selection	SMA	PSNR↑	LPIPS↓	CLIPIQA↑	MUSIQ↑
Full PixelDiT-SR	Full	×	25.543	**0.3714**	**0.7296**	**66.82**
Lightweight-Consecutive	Consecutive	×	**25.825**	0.4112	0.6407	61.87
Lightweight-Uniform	Uniform	×	25.739	0.3838	0.6144	63.86
LiPDiT-SR w/o SMA	MDBS	×	25.390	0.3798	0.6598	66.08
LiPDiT-SR	MDBS	✓	25.373	0.3800	0.6788	66.18

Effect of Each Component. Table 4 compares the selection strategies and SMA under identical budgets and training conditions. MDBS outperforms consecutive and uniform selection on all perceptual metrics, improving LPIPS from 0.4112/0.3838 to 0.3798, CLIPIQA from 0.6407/0.6144 to 0.6598, and MUSIQ from 61.87/63.86 to 66.08. This shows that the importance of a patch block is better characterized by its effect on pixel-level conditioning than by its depth or position. SMA further raises CLIPIQA and MUSIQ to 0.6788 and 66.18 with negligible changes in PSNR and LPIPS, narrowing the perceptual gap without affecting deployed inference.

Table 5: Comparison of different lightweight-model supervision targets on RealSR.

Target	PSNR↑	LPIPS↓	CLIPIQA↑	MUSIQ↑
Task loss only	25.390	0.3798	0.6598	66.08
Output KD	25.751	**0.3719**	0.6377	65.04
Semantic feature KD	**25.958**	0.3825	0.6363	63.37
SMA (ours)	25.373	0.3800	**0.6788**	**66.18**

Choice of Alignment Target. Table 5 compares four targets on the same lightweight architecture. Output KD and semantic feature KD attain the best LPIPS (0.3719) and PSNR (25.958), respectively, but reduce perceptual quality. SMA instead achieves the highest CLIPIQA (0.6788) and MUSIQ (66.18) while largely maintaining fidelity. Generic pixel or feature matching encourages conservative reconstruction; aligning the modulation interface leaves the preserved pixel pathway greater freedom to synthesize natural details. This fidelity–perception split supports interface-level supervision for preserving texture generation after pruning.

Table 6: Ablation of patch-level depth on RealSR. Inference time is measured end-to-end with batch size one.

Variant	Patch Blocks	PiT Blocks	Inference Time (ms)↓	PSNR↑	LPIPS↓	CLIPIQA↑	MUSIQ↑
Full PixelDiT-SR	26	4	22.67	**25.543**	**0.3714**	**0.7296**	**66.82**
LiPDiT-SR-22	22	4	17.17	25.389	0.3755	0.7064	66.82
LiPDiT-SR-18	18	4	15.06	25.373	0.3800	0.6788	66.18
LiPDiT-SR-14	14	4	12.76	25.359	0.3876	0.6280	64.53
LiPDiT-SR-10	10	4	**10.65**	25.156	0.4052	0.5473	60.73

Patch Budget and Path Allocation. Table 6 shows the quality–latency trade-off. Reducing 26 blocks to 22 lowers latency from 22.67 to 17.17 ms (24%) while retaining MUSIQ 66.82, with a 0.154 dB PSNR loss and CLIPIQA 0.7064 versus 0.7296. The 18-block model reaches 15.06 ms, with PSNR and LPIPS differing from the full model by 0.170 dB and 0.0086. Relative to 22 blocks, it saves another 12% latency for losses of 0.0276 CLIPIQA and 0.64 MUSIQ. At 14 blocks, CLIPIQA falls by 0.0508 to 0.6280 despite similar PSNR. Thus, 22 blocks favor quality, 18 offer our chosen speed–quality balance, and smaller budgets increasingly compromise perception.

5 Conclusion

We presented LiPDiT-SR, a VAE-free one-step SR model that compresses the cost-dominant patch-level pathway while preserving pixel-level rendering. MDBS selects blocks by their influence on downstream modulation, and SMA maintains this conditioning during progressive pruning without inference overhead. Experiments show competitive restoration quality with a 1.6× speedup, supporting function-aware compression of pixel diffusion Transformers.

The preserved pixel-level pathway limits further acceleration, and the fixed 18-block budget selected by offline calibration cannot adapt to image content or hardware. Progressive training also requires both the reference and current model, increasing training cost. Our evaluation covers $\times 4$ SR at a fixed resolution; robustness across scales and larger inputs remains untested. Future work will explore adaptive budgets and pixel-pathway compression.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript writing, implementation and editing of the method's code, and LaTeX formatting. The research ideas and original methodological contributions were developed by the authors without generative AI assistance. The authors reviewed the AI-assisted text and code and take full responsibility for the manuscript, implementation, and reported results.

REFERENCES

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 IEEE/CVF International Conference on Computer Vision*, pp. 3086–3095, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28208–28220, 2025. doi: 10.1109/CVPR52734.2025.02627.

Zhennan Chen, Junwei Zhu, Xu Chen, Jiangning Zhang, Xiaobin Hu, Hanzhen Zhao, Chengjie Wang, Jian Yang, and Ying Tai. DiP: Taming diffusion models in pixel space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36136–36146, 2026.

Katherine Crowson, Stefan Andreas Baumann, Alex Birch, Tanishq Mathew Abraham, Daniel Z. Kaplan, and Enrico Shippole. Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9550–9575, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23174–23184, 2025.

Linwei Dong, Qingnan Fan, Yuhang Yu, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. TinySR: Shallow diffusion transformers for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 5075–5085, 2026.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021.

Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. SRDiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. doi: 10.1016/j.neucom.2022.01.029.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. DiffBIR: Toward blind image restoration with generative diffusion prior. In *Computer Vision – ECCV 2024*, pp. 430–448, 2024. doi: 10.1007/978-3-031-73202-7_25.

Jian Ma, Qirong Peng, Xujie Zhu, Peixing Xie, Chen Chen, and Haonan Lu. Pluggable pruning with contiguous layer distillation for diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18640–18650, 2026a.

Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. DeCo: Frequency-decoupled pixel diffusion for end-to-end image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43600–43610, 2026b.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2023. doi: 10.1109/TPAMI.2022.3204461.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2333–2343, 2025.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023. doi: 10.1609/aaai.v37i2.25353.

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*, 132(12):5929–5949, 2024a. doi: 10.1007/s11263-024-02168-7.

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 IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1905–1914, 2021.

Yan Wang, Shijie Zhao, Junlin Li, and Li Zhang. Eliminating VAE for fast and high-resolution generative detail restoration. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.10630.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25796–25805, 2024b.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

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*, volume 37, pp. 92529–92553, 2024a. doi: 10.52202/079017-2938.

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 IEEE/CVF 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022. doi: 10.1109/CVPRW56347.2022.00126.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *Computer Vision – ECCV 2024*, pp. 74–91, 2024. doi: 10.1007/978-3-031-73247-8_5.

Haoran You, Connelly Barnes, Yuqian Zhou, Yan Kang, Zhenbang Du, Wei Zhou, Lingzhi Zhang, Yotam Nitzan, Xiaoyang Liu, Zhe Lin, Eli Shechtman, Sohrab Amirghodsi, and Yingyan Celine Lin. Layer- and timestep-adaptive differentiable token compression ratios for efficient diffusion transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18072–18082, 2025a.

Weiyi You, Mingyang Zhang, Leheng Zhang, Xingyu Zhou, Kexuan Shi, and Shuhang Gu. Consistency trajectory matching for one-step generative super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12747–12756, 2025b.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25669–25680, 2024.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. PixelDiT: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14273–14282, 2026.

Zhihang Yuan, Hanling Zhang, Lu Pu, Xuefei Ning, Linfeng Zhang, Tianchen Zhao, Shengen Yan, Guohao Dai, and Yu Wang. DiTFastAttn: Attention compression for diffusion transformer models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. ResShift: Efficient diffusion model for image super-resolution by residual shifting. In *Advances in Neural Information Processing Systems*, volume 36, pp. 13294–13307, 2023.

Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23153–23163, 2025. doi: 10.1109/CVPR52734.2025.02156.

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 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.