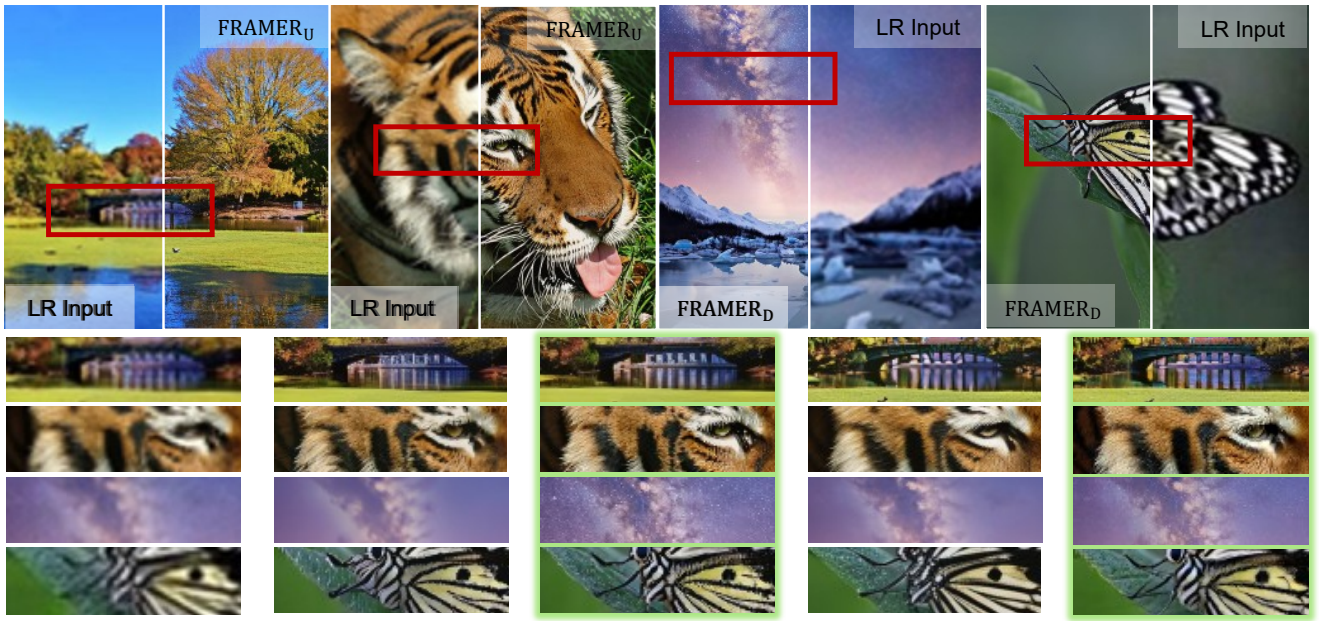
FRAMER: Frequency-Aligned Self-Distillation with Adaptive Modulation Leveraging Diffusion Priors for Real-World Image Super-Resolution

Seungho Choi, Jeahun Sung, and Jihyong Oh†
Chung-Ang University
Seoul, Republic of Korea
choiseungho1019@gmail.com, {jhseong, jihyongoh}@cau.ac.kr
https://cmlab-korea.github.io/FRAMER/



LR Input | DiT4SR [10] | FRAMER$_D$ | PiSA-SR [39] | FRAMER$_U$

Figure 1. **Qualitative comparison with recent Real-ISR methods on real-world images.** Our FRAMER models produce sharper edges and richer details, leading to more visually natural and faithful restoration results. More qualitative results are provided in Supplementary Sec. C.

Abstract

*Real-image super-resolution (Real-ISR) seeks to recover HR images from LR inputs with mixed, unknown degradations. While diffusion models surpass GANs in perceptual quality, they under-reconstruct high-frequency (HF) details due to a low-frequency (LF) bias and a depth-wise "low-first, high-later" hierarchy. We introduce **FRAMER**, a plug-and-play training scheme that exploits diffusion priors without changing the backbone or inference. At each denoising step, the final-layer feature map teaches all intermediate layers. Teacher and student feature maps are decomposed into LF/HF bands via FFT masks to align supervision with the model's internal frequency hierarchy. For LF, an Intra Contrastive Loss (IntraCL) stabilizes globally shared structure. For HF, an Inter Contrastive Loss (InterCL) sharpens instance-specific details using random-layer and in-batch negatives. Two adaptive modulators, Frequency-based Adaptive Weight (FAW) and Frequency-based Alignment Modulation (FAM), reweight per-layer LF/HF signals and gate distillation by current similarity. Across U-Net and DiT backbones (e.g., Stable Diffusion 2, 3), FRAMER consistently improves PSNR/SSIM and perceptual metrics (LPIPS, NIQE, MANIQA, MUSIQ). Ablations validate the final-layer teacher and random-layer negatives.*

†Corresponding author

1. Introduction

Real Image Super-Resolution (Real-ISR) [43] aims to recover a high-resolution (HR) image from a low-resolution (LR) counterpart suffering from mixed, unknown degradations (e.g., blur, noise) [43]. Unlike traditional SR [9], Real-ISR must simultaneously remove complex degradations and generate realistic details, demanding stronger model capacity and robustness [26, 48].

While SR methods evolved from early CNNs to GANs [13], GANs often suffer from training instability under the complex degradations of Real-ISR [22, 38]. Diffusion models have surpassed them by offering stable training and superior perceptual quality [8, 36]. Leveraging large-scale pre-trained text-to-image (T2I) models (e.g., Stable Diffusion with U-Net [35] or DiT [31] backbones) is a promising direction due to their rich, real-world priors [31, 32, 52]. However, these priors are optimized for creative generation, not the high-fidelity, task-constrained demands of Real-ISR [42, 48]. The key challenge is adapting these priors for our task.

Despite their success, diffusion models still struggle to reconstruct fine high-frequency (HF) details, often yielding over-smoothed results [29]. We trace this to a fundamental low-frequency (LF) bias stemming from two properties. First, natural image frequency distributions are inherently LF-dominant, an imbalance that is severely exacerbated in the LR inputs (Fig. 2). The standard noise-prediction loss [16] thus favors these dominant LF components to reduce overall loss, inevitably undertraining HF signals [7, 12]. Second, we identify a depth-wise "low-first, high-later" hierarchy: an analysis of layer-wise feature maps (Fig. 3) shows that LF features stabilize early in the network, while HF features converge only near the final layers. A conventional, frequency-agnostic loss is fundamentally misaligned with this hierarchy. It supplies redundant LF-biased gradients to early layers (where LF has already stabilized) while starving the later, HF-refining layers of the necessary optimization signals.

A simple approach, adding explicit frequency-domain losses, fails due to a domain mismatch between high-dimensional features and pixel-space targets [40, 53]. Self-distillation (SD) [15], however, offers a well-aligned alternative [5, 17]. We treat the final-layer feature map as the *teacher* and intermediate layers as *students*. This avoids domain mismatch, as both operate within the same feature space, providing a consistent and domain-compatible distillation signal [51].

However, standard SD is frequency-agnostic. To correct the identified spectral bias and layer-wise misalignment, we propose **Fr**equency-**A**ligned Self-Distillation with Adaptive **M**odulation Lev**er**aging Diffusion Priors, namely **FRAMER**, a layer-adaptive SD framework with four frequency-aware components. First, **Intra Contrastive Loss (IntraCL)** (Sec. 3.2) focuses on LF stability. Second, **Inter Contrastive Loss (InterCL)** (Sec. 3.3) targets HF detail sharpness. Third, **Frequency-based Adaptive Weight (FAW)** (Sec. 3.4) adaptively modulates distillation strength across layers and frequencies. Finally, **Frequency-based Alignment Modulation (FAM)** (Sec. 3.5) controls strength based on student-teacher similarity to prevent collapse and stabilize training.

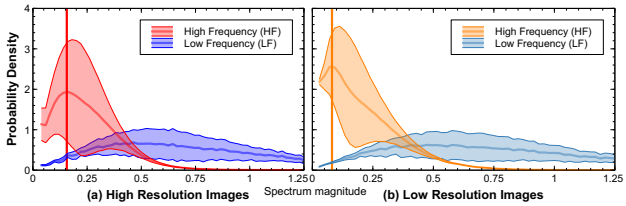


Figure 2. **Band-wise magnitude densities with shared bins.** For each feature map, we compute the 2D FFT and collect magnitudes $|F|$ within LF and HF rings. We plot mean $\pm$ σ densities over samples for $\log(1+|F|)$ using common bin edges (HF: red or yellow, LF: blues). LF magnitudes span a broader and heavier range, whereas HF magnitudes concentrate narrowly near small values, indicating LF dominance that biases unified training toward LF and undertrains HF details. All statistics are computed on the 100-image DIV2K [1] test set. Densities integrate to 1; any right-edge spike is due to percentile clipping used only for visualization.

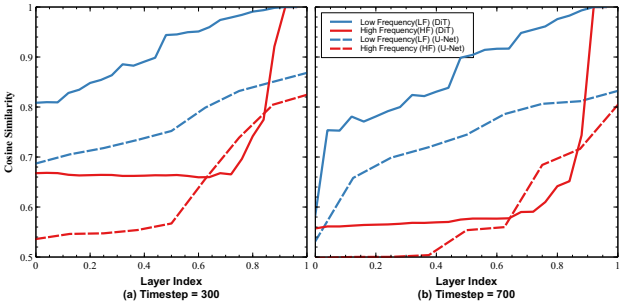


Figure 3. **Layer-wise cosine similarity of LF and HF feature maps in U-Net [35] (dotted line) and DiT [31] (solid line).** (a) low-noise timestep (t=300), (b) high-noise timestep (t=700).Using the final-layer feature map as reference, LF similarity converges faster in earlier layers, whereas HF similarity rises abruptly in later layers. This reveals a "low-first, high-later" depth-wise hierarchy (i.e., an LF bias), motivating our layer-adaptive, frequency-aware training strategy. (For comparability, layer depth is normalized to [0, 1].)

In summary, our main contributions are:

- From an SR training perspective, we analyze empirical evidence of a layer-wise "low-first, high-later" frequency hierarchy in diffusion models (Fig. 3) and identify it as a key optimization challenge for Real-ISR.
- We propose **FRAMER**, a self-distillation framework that introduces frequency-aligned contrastive losses, IntraCL

for LF stability and InterCL for HF details, to counteract the LF bias in SR training tendency of the standard noise-prediction loss [16].
- We further introduce adaptive distillation mechanisms, FAW and FAM, based on the internal frequency hierarchy. Together they suppress unstable gradients, prevent early-layer collapse and overfitting, accelerate convergence, resulting in improving SR quality in terms of diverse metrics with better training stability.
- FRAMER is a plug-and-play training technique, effectively leveraging the rich priors of pre-trained text-to-image models, such as Stable Diffusion 2 [32] (U-Net) and Stable Diffusion 3 [11] (DiT), without altering their architecture or inference.

2. Related Work

Image Super-Resolution. Early SR followed supervised regression under synthetic bicubic downsampling, then moved to blind SR for unknown degradations, and progressed to Real-ISR emphasizing realistic evaluation [43]. CNN-based methods improved distortion but were often oversmoothed; GANs increased realism but suffered from stability issues. Diffusion models advanced SR with stronger perceptual quality [8, 36]. Research then adopted pre-trained models like Stable Diffusion to leverage real-world priors [6, 10, 39]. However, these pipelines supervise all layers and frequencies with a single, standard noise-prediction loss, and thus do not exploit the early LF and late HF behavior along timesteps and depth. FRAMER resolves this gap with a novel self-distillation framework that is both frequency-aligned and layer-adaptive.

Frequency-Aware Learning in Diffusion Models. Several works shape the diffusion process to be frequency-aware. Approaches include manipulating forward noising in the frequency domain [19], designing frequency-guided multi-scale diffusion with wavelets [44], introducing HF-preserving objectives to align features [37], or using training-free inference-time modulation [23]. These methods increase frequency awareness but generally rely on fixed schedules or auxiliary modules. Critically, they do not adapt supervision to what each layer actually changes at a given timestep. FRAMER addresses this specific gap with its FAW, which modulates supervision based on the actual layer-wise change rate, and FAM, which gates distillation based on student-teacher alignment, rather than relying on a pre-defined, fixed schedule.

Self-Distillation in Diffusion Models. Self-distillation (SD) has been applied to improve generative fidelity beyond sampling acceleration. Methods include unified SD losses [47], step-level fusion aggregating predictions across timesteps [28], and representation alignment enforcing feature consistency [3, 18]. These techniques are powerful for aligning representations but remain fundamentally frequency-agnostic; they align the entire feature map, thus implicitly inheriting the LF-bias of standard losses. FRAMER complements this line by explicitly introducing frequency-awareness to the SD framework. By using a dedicated IntraCL for LF stability and an InterCL for HF detail, it separates the distillation objectives based on spectral roles, directly counteracting the LF-bias instead of inheriting it.

3. Proposed Method

Training diffusion models for Real-ISR faces two key challenges [7, 12]: (1) the noise-prediction loss is biased toward LF components, and (2) this loss is misaligned with the network's "low-first, high-later" processing, leaving HF layers under-optimized (Fig. 3).

Simple auxiliary frequency-based losses for intermediate layers fail, introducing a cross-domain objective mismatch [40, 53] that forces high-level features to regress to different-domain targets, destabilizing training [34].

To avoid this mismatch, we leverage priors from pre-trained diffusion backbones (e.g., SD2, SD3) [11, 32], which encode rich natural-image statistics. We realign these priors for Real-ISR via frequency-aware, layer-adaptive supervision without changing inference. We use self-distillation (SD) [5, 15, 17]: intermediate layers (*students*) are supervised by the final-layer feature map (*teacher*) [16, 31, 32]. This provides a domain-compatible signal, avoiding the cross-domain mismatch.

3.1. Observations

(1) LF-dominant training statistics in Real-ISR. In Real-ISR, HR/LR images have unbalanced band-wise magnitudes: LF is broad and heavy, while HF is narrow and small. Thus, the noise-prediction objective is LF-biased and under-trains HF [7, 12] (cf. Fig. 2).

(2) Shared LF vs. instance-specific HF. Per Fig. 5, LF features show high cross-sample similarity (shared structures), while HF features show low similarity (instance-specific details). This implies in-batch negatives are false negatives for LF but informative for HF.

(3) Depth-wise "low-first, high-later" hierarchy. Consistent with Fig. 3, diffusion backbones show a "low-first, high-later" frequency-progressive hierarchy: LF stabilizes in early layers, while HF similarity rises in later layers.

(4) Degrees of alignment increase along layer depth. Per Fig. 3, student–teacher feature alignment is low in early layers and increases with depth.

Overview. Building on Sec. 3.1, we propose **FRAMER** (**Fr**equency-**A**ligned Self-Distillation with Adaptive **M**odulation Lev**er**aging Diffusion Priors) (Fig. 4). During training, we create LR inputs I_{LR} from HR images R via random degradation [43] and get captions via LLaVA [27]. The model is conditioned on $(I_{\text{LR}}, Z_T, \text{caption})$. FRAMER

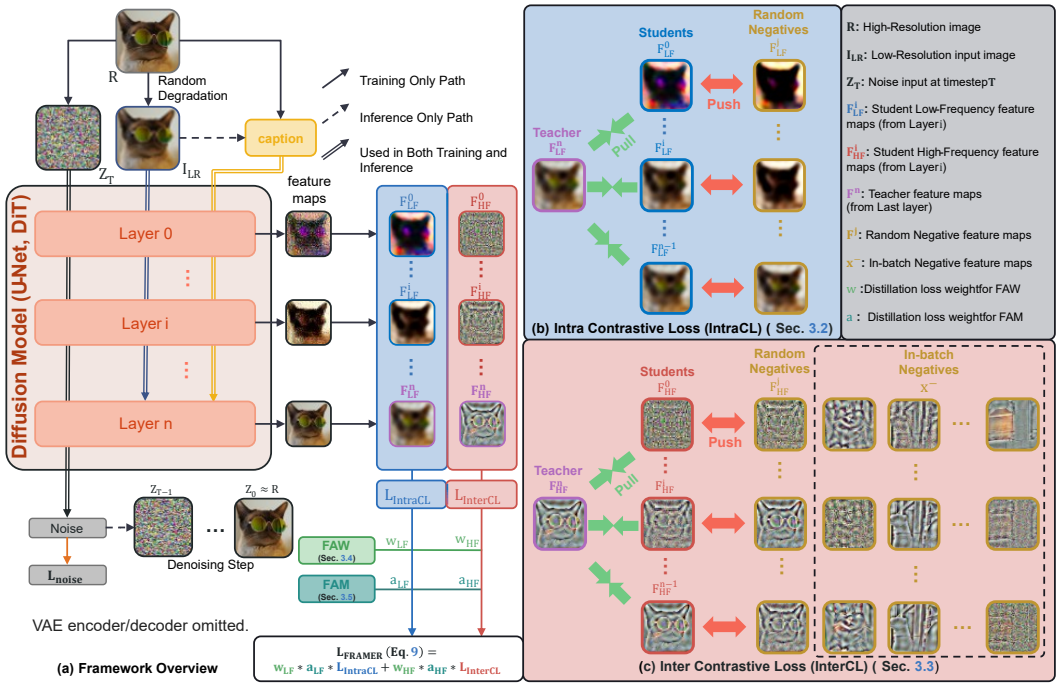


Figure 4. **FRAMER: Frequency-Aligned Self-Distillation with Adaptive Modulation Leveraging Diffusion Priors** (inspired by Sec. 3.1). **(a) Framework Overview.** During training, from an High-Resolution image R, we create I_{LR} by random degradation [43], downsampling, and resizing back to the size of R. We use LLaVA [27] to generate a caption. The diffusion backbone (U-Net [35]/DiT [31]) takes I_{LR}, noise Z_T, and the caption as inputs; FRAMER is applied only during training and jointly uses IntraCL (Sec. 3.2) and InterCL (Sec. 3.3), modulated by FAW (Sec. 3.4) and FAM (Sec. 3.5). FRAMER is a training framework that adds auxiliary loss components only during training. At inference, it uses the original backbone without any modification, making it fully plug-and-play. **(b) Intra Contrastive Loss.** Within a single image, pull $F_{\mathrm{LF}}^{(i)}$ toward $F_{\mathrm{LF}}^{(n)}$ and push it away from a randomly sampled same-image layer $F_{\mathrm{LF}}^{(j)}$ (no in-batch negatives). **(c) Inter Contrastive Loss.** Attract $F_{\mathrm{HF}}^{(i)}$ to $F_{\mathrm{HF}}^{(n)}$ and repel a random-layer negative $F_{\mathrm{HF}}^{(j)}$ (same image) and in-batch negatives x^- (other images).

is a training-only, plug-and-play strategy with no inference overhead. During training, the final-layer feature map (*teacher*) supervises intermediate layers (*students*). We decompose teacher/student features into LF/HF bands (see Sec. 5.1, Sec. 5.2 for justification). We then: (i) apply IntraCL (Sec. 3.2) to the LF band to stabilize shared structure (no in-batch negatives); (ii) apply InterCL (Sec. 3.3) to the HF band to sharpen instance-details (using in-batch/random-layer negatives); (iii) compute FAW (Sec. 3.4) to adaptively weight distillation based on student–teacher discrepancy; and (iv) apply FAM (Sec. 3.5) to gate the distillation by student–teacher alignment. These terms are added to the noise-prediction loss. The full algorithm is provided in Supplementary Sec. A.1.

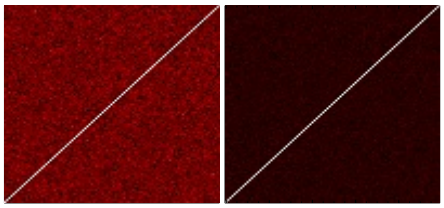
3.2. Intra Contrastive Loss (IntraCL) for LF

As observed in Sec. 3.1 (1, 2) and visualized in Fig. 5, LF feature maps are highly shared across training samples, whereas HF feature maps are instance-specific. Consequently, using in-batch negatives for LF easily introduces false negatives because many images share similar coarse structures. We therefore introduce an LF-oriented IntraCL that compares a student only against its teacher and a randomly sampled layer within the same network. This stabilizes globally shared structures and promotes layer-wise refinement of LF representations.

Concretely, let $\mathbf{F}^{(i)}$ denote the feature map from an intermediate layer i of the diffusion backbone (U-Net or DiT). We obtain the LF representation $\mathbf{F}_{\mathrm{LF}}^{(i)}$ by frequency-decomposing $\mathbf{F}^{(i)}$ via 2D FFT and applying a predefined LF mask to its magnitude spectrum. Let $\mathbf{F}_{\mathrm{LF}}^{(n)}$ denote the LF representation of the final-layer feature map (the teacher), and let $\mathbf{F}_{\mathrm{LF}}^{(j)}$ be the LF representation of a randomly sampled student layer j.

We define the similarity function $\mathrm{sim}(\cdot,\cdot)$ as the cosine similarity. This is computed as the dot product between L2-normalized feature maps $\hat{\mathbf{F}} = \mathbf{F}/\|\mathbf{F}\|_2$. The similarity scores are $s_{+,\mathrm{LF}}^{(i)} = \mathrm{sim}(\hat{\mathbf{F}}_{\mathrm{LF}}^{(i)}, \hat{\mathbf{F}}_{\mathrm{LF}}^{(n)})$; $s_{-,\mathrm{LF}}^{(i)} = \mathrm{sim}(\hat{\mathbf{F}}_{\mathrm{LF}}^{(i)}, \hat{\mathbf{F}}_{\mathrm{LF}}^{(j)})$, where j is sampled uniformly at each step from $\mathcal{J}_i = \{1,\ldots,n\} \setminus \{i,n\}$, $j \sim \mathrm{Unif}(\mathcal{J}_i)$.

The IntraCL is formulated as a log-softmax over these



(a) Low-frequency similarity (b) High-frequency similarity

Figure 5. **Visualization of feature maps similarity matrices across training samples in different frequencies** (brighter/redder indicates higher similarity). (a) LF exhibits strong cross sample similarity, reflecting shared structural information and motivating the use of **IntraCL** (Sec. 3.2) for stabilizing global structure learning. (b) HF shows weak cross sample similarity and strong sample specific variation, justifying the use of **InterCL** (Sec. 3.3) to promote fine grained, instance level discrimination without over sharpening. Detailed descriptions of the training samples are in Sec. 4.1

scores:

$$\mathcal{L}_{\text{IntraCL}}^{(i)} = -\log \frac{\exp\left(s_{+,\text{LF}}^{(i)}\right)}{\exp\left(s_{+,\text{LF}}^{(i)}\right) + \exp\left(s_{-,\text{LF}}^{(i)}\right)}. \quad (1)$$

This self-distillation objective pulls the student toward the teacher's LF representation. Because the teacher and students operate within the same high-dimensional feature space, this approach avoids the domain mismatch problem and provides a consistent and domain-compatible distillation signal to stabilize the network's understanding of global structures.

3.3. Inter Contrastive Loss (InterCL) for HF

As observed in Sec. 3.1 (1, 2) and confirmed by the cross-sample statistics in Fig. 5, HF components encode fine-grained, sample-specific details and show low cross-sample similarity, which makes in-batch negatives informative and valid. Accordingly, we introduce an HF-oriented InterCL that attracts a student's HF feature map to the teacher's HF feature map while repelling it from two types of negatives: (i) a random-layer HF feature map within the same image (to enforce layer-wise progression) and (ii) *in-batch* HF feature maps from other images (to encourage instance discrimination). This objective directly counteracts the LF bias and delivers targeted optimization signals to the slowly converging HF components, sharpening image-specific details.

Analogous to the LF branch, we obtain the HF representation $\mathbf{F}_{\text{HF}}^{(i)}$ by applying the predefined HF mask to the 2D FFT magnitude spectrum of the intermediate-layer feature map $\mathbf{F}^{(i)}$. Using the L2-normalized feature maps $\hat{\mathbf{F}}_{\text{HF}}$, we define the positive and random-layer negative similarity scores: $s_{+,\text{HF}}^{(i)} = \text{sim}(\hat{\mathbf{F}}_{\text{HF}}^{(i)}, \hat{\mathbf{F}}_{\text{HF}}^{(n)})$; $s_{-,\text{HF}}^{(i)} = \text{sim}(\hat{\mathbf{F}}_{\text{HF}}^{(i)}, \hat{\mathbf{F}}_{\text{HF}}^{(j)})$, where j is sampled uniformly from $\mathcal{J}_i = \{1, \ldots, n\} \setminus \{i, n\}$.

Let x^- denote the L2-normalized HF feature maps from other training samples in the current batch. We aggregate these in-batch negatives as: $S_{\text{neg}}^{(i)} = \sum_{x^-} \exp\big(\text{sim}(\hat{\mathbf{F}}_{\text{HF}}^{(i)}, x^-)\big)$.

The InterCL loss is then formulated as a log-softmax over the positive, the random-layer negative, and the *in-batch* negatives. The HF branch and its InterCL objective are depicted in Fig. 4(c).

$$\mathcal{L}_{\text{InterCL}}^{(i)} = -\log \frac{\exp\left(s_{+,\text{HF}}^{(i)}\right)}{\exp\left(s_{+,\text{HF}}^{(i)}\right) + \exp\left(s_{-,\text{HF}}^{(i)}\right) + S_{\text{neg}}^{(i)}}. \quad (2)$$

Here, the random-layer negative j enforces layer-wise progression, while the in-batch term $S_{\text{neg}}^{(i)}$ encourages instance-aware discrimination. This objective directly counteracts the LF bias and provides targeted optimization signals to the slowly converging HF components, as shown in Fig. 3, helping the model reconstruct fine, HF details.

3.4. Frequency-based Adaptive Weight (FAW)

As observed in Sec. 3.1 (3) and Fig. 3, diffusion backbones follow a depth-wise "low-first, high-later" frequency hierarchy: early layers prioritize stabilizing LF representations, whereas HF similarity emerges chiefly in later layers. To align the self-distillation losses (IntraCL, InterCL) with this hierarchy, we introduce Frequency-based Adaptive Weight (FAW), which decomposes self-distillation across layer depth and frequency bands and, at each layer, sets the loss weight for each band according to its relative difference to the final layer. This yields normalized per-layer LF/HF weights that directly scale the corresponding distillation losses. The effectiveness of this FAW mechanism is empirically validated in Sec. 5.4

We use the same predefined binary masks M_{LF} and M_{HF} from Sections 3.2 and 3.3. Let $\left|\mathbf{F}^{(i)}\right|$ be the FFT magnitude at layer i. The per-frequency average magnitudes are:

$$E_{\text{LF}}^{(i)} = \frac{\sum\left(\left|\mathbf{F}^{(i)}\right| \odot M_{\text{LF}}\right)}{\sum M_{\text{LF}} + \varepsilon}, \quad E_{\text{HF}}^{(i)} = \frac{\sum\left(\left|\mathbf{F}^{(i)}\right| \odot M_{\text{HF}}\right)}{\sum M_{\text{HF}} + \varepsilon}. \quad (3)$$

Here, n denotes the index of the final (teacher) layer and ε a small constant (e.g., 10^{-9}) for numerical stability; we then calculate the relative difference to the final layer n:

$$\Delta_{\text{LF}}^{(i)} = \frac{\left|E_{\text{LF}}^{(n)} - E_{\text{LF}}^{(i)}\right|}{E_{\text{LF}}^{(i)} + \varepsilon}, \quad \Delta_{\text{HF}}^{(i)} = \frac{\left|E_{\text{HF}}^{(n)} - E_{\text{HF}}^{(i)}\right|}{E_{\text{HF}}^{(i)} + \varepsilon}. \quad (4)$$

Finally, the weights $\mathbf{w}^{(i)}$ are normalized via softmax to sum to one, which mitigates scale-induced spectral bias:

$$\mathbf{w}^{(i)} = \text{softmax}\Big([\Delta_{\text{LF}}^{(i)}, \Delta_{\text{HF}}^{(i)}]\Big), \quad w_{\text{LF}}^{(i)} + w_{\text{HF}}^{(i)} = 1. \quad (5)$$

Given the normalized, per-layer weights $\mathbf{w}^{(i)}$ in Eq. 5, we combine the LF/HF contrastive objectives into a single FAW-only self-distillation term by linearly weighting the branches as follows:

$$\mathcal{L}^{(i)}_{\text{FRAMER}} = w^{(i)}_{\text{LF}}\,\mathcal{L}^{(i)}_{\text{IntraCL}} + w^{(i)}_{\text{HF}}\,\mathcal{L}^{(i)}_{\text{InterCL}}. \tag{6}$$

3.5. Frequency-based Alignment Modulation (FAM)

As observed in Sec. 3.1 (4) and Fig. 3, the student–teacher alignment increases along layer depth: in early layers, student features are far from the final-layer teacher, so self-distillation from earlier layers can destabilize optimization, whereas those from later layers are more helpful. We therefore propose Frequency-based Alignment Modulation (FAM), which gates the self-distillation strength by the current LF/HF alignment, suppressing it when alignment strength is low and adaptively increasing as a degree of alignment improves. The effectiveness of this FAM mechanism is empirically validated in Sec. 5.4.

We first define the alignment scores $a^{(i)}$ as the cosine similarity between the L2-normalized student and teacher feature maps (using $\hat{\mathbf{F}}$ defined in Sec. 3.2), passed through a ReLU to ensure non-negativity:

$$a^{(i)}_{\text{LF}} = \text{ReLU}\big(\text{sim}(\hat{\mathbf{F}}^{(i)}_{\text{LF}}, \hat{\mathbf{F}}^{(n)}_{\text{LF}})\big), \quad a^{(i)}_{\text{HF}} = \text{ReLU}\big(\text{sim}(\hat{\mathbf{F}}^{(i)}_{\text{HF}}, \hat{\mathbf{F}}^{(n)}_{\text{HF}})\big). \tag{7}$$

We then use these alignment scores, with detached gradients, to gate the FAW weights:

$$\tilde{w}^{(i)}_{\text{LF}} = w^{(i)}_{\text{LF}} \cdot \text{stopgrad}\big(a^{(i)}_{\text{LF}}\big), \quad \tilde{w}^{(i)}_{\text{HF}} = w^{(i)}_{\text{HF}} \cdot \text{stopgrad}\big(a^{(i)}_{\text{HF}}\big). \tag{8}$$

The final, layer-wise distillation loss for our FRAMER framework is the weighted sum of the two contrastive losses:

$$\mathcal{L}^{(i)}_{\text{FRAMER}} = \tilde{w}^{(i)}_{\text{LF}}\;\mathcal{L}^{(i)}_{\text{IntraCL}} + \tilde{w}^{(i)}_{\text{HF}}\;\mathcal{L}^{(i)}_{\text{InterCL}}. \tag{9}$$

Total objective. The final training objective combines the standard noise-prediction loss ($\mathcal{L}_{\text{noise}}$) with the sum of our layer-wise, frequency-gated distillation terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{noise}} + \sum_{i=1}^{N} \mathcal{L}^{(i)}_{\text{FRAMER}}. \tag{10}$$

This full objective aligns self-distillation with the internal frequency hierarchy, delivering targeted frequency-aware signals that counteract LF bias [7, 12] and, via FAW and FAM, stabilize optimization by gating distillation to layer-wise discrepancy and alignment so that early layers are stabilized and HF details in later layers are sharpened while balancing globally coherent structure with instance-specific detail. All auxiliary heads and losses are strictly training-only and removed at test time with no architectural or inference changes, making FRAMER a plug-and-play training method that leverages the strong priors of large pre-trained T2I diffusion backbones such as Stable Diffusion 2 [32] (U-Net [35]) and SD3 [11] (DiT [31]) to improve perceptual realism and high-frequency fidelity.

4. Experiments

4.1. Experimental Settings

Datasets. To ensure fair comparisons, we follow the same experimental setup as SeeSR [48] and DiT4SR [10]. Our training set is composed of a mixture of images from DIV2K [1], DIV8K [14], Flickr2K [41], and the initial 10K face images from FFHQ [20]. The Real-ESRGAN [43] degradation pipeline for SR training is adopted to generate LR–HR training pairs using the same parameter configuration as SeeSR. The input LR and HR resolutions are set to 128×128 and 512×512, respectively. For evaluation, we employ four real-world datasets: DrealSR [46], RealSR [4], RealLR200 [48], and RealLQ250 [2]. All experiments are conducted with an upscaling factor of ×4.

Metrics. To ensure comparability with prior works, we report both full-reference and no-reference metrics. Specifically, we include PSNR and SSIM [45] as full-reference measures for distortion/error, while acknowledging their limited alignment with human perception in restoration tasks; for perceptual fidelity we adopt LPIPS [52]. For no-reference quality assessment, we use NIQE [30] as an opinion-unaware metric, and MANIQA [49] and MUSIQ [21] as recent learning-based NR-IQA measures.

Implementation Details. All implementation details, including loss functions, frequency decomposition, training iterations, etc., are provided in the *Suppl.* for reproducibility.

4.2. Comparison with Other Methods

We compare our method with state-of-the-art Real-ISR methods, including Swin Transformer-based methods, U-Net-based methods, and DiT-based methods. Among these, our FRAMER is instantiated on the U-Net-based PiSA-SR and the DiT-based DiT4SR to validate its architectural flexibility. FRAMER is trained in a fully plug-and-play manner under exactly the same settings as each corresponding baseline, without any architecture-specific tuning. Unlike DiT, whose feature resolutions remain identical across layers, U-Net changes the feature resolution between stages; therefore, we insert a 1×1 convolution and a resize operation to align dimensions when integrating FRAMER into the U-Net, as consistent feature shapes are required for self-distillation.

Quantitative Comparisons. We conduct quantitative evaluations of our proposed FRAMER framework against recent state-of-the-art Real-ISR methods across four real-

Table 1. **Quantitative comparison of real-world image super-resolution methods.** We evaluate both fidelity metrics (PSNR↑, SSIM↑, LPIPS↓) and perceptual quality metrics (NIQE↓, MANIQA↑, MUSIQ↑). Methods are grouped by architecture type (Swin-based, U-Net-based, DiT-based). **Best** and Second best results are highlighted. The green percentages for our FRAMER models indicate the relative improvement over their respective baselines, PiSA-SR (for FRAMER_{U}) and DiT4SR (for FRAMER_{D}).

Datasets	Metrics	Swin-based		U-Net-based			DiT-based		
		SwinIR [24] (ICCV'21)	ResShift [50] (NeurIPS'23)	SeeSR [48] (CVPR'24)	PiSA-SR [39] (CVPR'25)	FRAMER_{U} (**Ours**)	DreamClear [2] (NeurIPS'24)	DiT4SR [10] (ICCV'25)	FRAMER_{D} (**Ours**)
DrealSR	PSNR ↑	25.51	25.97	25.92	26.18	**26.96** (+3.0%)	23.41	23.64	24.73 (+4.6%)
	SSIM ↑	0.742	0.704	0.734	0.752	**0.786** (+4.5%)	0.652	0.640	0.687 (+7.3%)
	LPIPS ↓	0.366	0.520	0.395	0.368	**0.333** (+9.5%)	0.493	0.442	0.412 (+6.8%)
	NIQE ↓	6.555	8.763	6.452	6.136	**5.386** (+12.2%)	6.481	6.780	5.959 (+12.1%)
	MANIQA ↑	0.330	0.341	0.507	0.490	**0.595** (+21.4%)	0.418	0.441	0.514 (+16.6%)
	MUSIQ ↑	55.26	56.17	64.99	66.15	**74.53** (+12.7%)	57.68	64.93	68.47 (+5.5%)
RealSR	PSNR ↑	23.56	23.51	23.69	24.02	**24.81** (+3.3%)	21.67	21.94	23.23 (+5.9%)
	SSIM ↑	0.719	0.697	0.700	0.719	**0.746** (+3.8%)	0.698	0.640	0.679 (+6.1%)
	LPIPS ↓	0.378	0.497	0.386	0.355	**0.328** (+7.6%)	0.489	0.414	0.371 (+10.4%)
	NIQE ↓	5.605	7.174	5.337	5.902	5.513 (+6.6%)	6.691	6.262	**5.083** (+18.8%)
	MANIQA ↑	0.359	0.338	0.542	0.412	0.484 (+17.5%)	0.436	0.459	**0.564** (+22.9%)
	MUSIQ ↑	61.47	60.67	69.80	68.20	70.03 (+2.7%)	64.89	67.67	**72.24** (+6.8%)
RealLR200	NIQE ↓	3.693	5.305	3.919	4.410	4.381 (+0.7%)	**3.229**	4.342	4.038 (+7.0%)
	MANIQA ↑	0.343	0.361	0.472	0.499	0.525 (+5.2%)	0.475	0.514	**0.552** (+7.4%)
	MUSIQ ↑	64.59	61.24	71.63	71.95	73.38 (+2.0%)	67.63	72.29	**74.30** (+2.8%)
RealLQ250	NIQE ↓	3.660	5.274	3.896	4.138	4.083 (+1.3%)	**3.503**	4.199	3.907 (+7.0%)
	MANIQA ↑	0.357	0.394	0.490	0.471	0.516 (+9.6%)	0.427	0.507	**0.546** (+7.7%)
	MUSIQ ↑	65.09	62.87	72.19	71.19	73.12 (+2.7%)	68.40	72.19	**73.91** (+2.4%)

world benchmarks, as shown in Table 1. On the DrealSR and RealSR datasets, FRAMER achieves performance on par with or superior to advanced approaches such as SeeSR and DiffBIR, demonstrating strong fidelity and perceptual quality. On the more challenging RealLR200 and RealLQ250 datasets, our method exhibits clear superiority, ranking first across all perceptual quality metrics. These results collectively highlight the effectiveness of FRAMER in producing high-quality and visually realistic restorations, benefiting from its robust generative prior and architecture-agnostic design.

Qualitative Comparisons. The qualitative results compared with recent Real-ISR methods are illustrated in Fig. 1. As shown, our method produces clearer textures and more faithful structures than existing approaches, particularly under severe degradations. Both FRAMER_{U} and FRAMER_{D} effectively recover fine details while preserving natural appearance, highlighting the strong restoration capability of our framework. More comprehensive qualitative comparisons can be found in the *Supplementary Material*.

5. Ablation Study

To validate the effectiveness of our proposed FRAMER, we conduct comprehensive ablations on its key components: (1) the frequency-specific contrastive distillation objective, (2) the design of IntraCL and InterCL, and (3) the adaptive mechanisms FAW and FAM. Unless otherwise noted, all experiments share the same setting: training is performed under the same conditions as Sec. 4.1, and evaluation is conducted on the RealSR dataset.

Table 2. **Ablation on the distillation objective.** Metrics are on RealSR. The **best** per column is in **bold**.

Method	Distillation Setup		Quality Metrics		Perceptual Metrics	
	Freq. Decomp.	Loss Type	PSNR↑	LPIPS↓	MUSIQ↑	MANIQA↑
`Baseline`	×	×	21.94	0.41	67.67	0.459
`+MSE Distill.`	×	MSE	22.36	0.41	68.11	0.467
`+MSE-Freq Distill.`	✓	MSE	22.59	0.39	65.74	0.417
`+CL-Freq Distill.`	✓	CL	**23.04**	**0.38**	**69.01**	**0.533**

5.1. Effectiveness of the Distillation Objective

As shown in Table 2, we first test the hypothesis that a frequency-specific contrastive objective is preferable to L2-based self-distillation for Real-ISR. Adding frequency-agnostic L2 between the teacher and students ("`+MSE Distill.`") yields modest gains over "`Baseline`" but does not correct the LF bias of the standard noise-prediction loss [16]. Introducing frequency decomposition with L2 ("`+MSE-Freq Distill.`") improves distortion metrics yet harms perceptual metrics, consistent with domain mismatch when regressing high-dimensional features to pixel/Fourier targets [40, 53]. In contrast, our frequency-specific contrastive variant ("`+CL-Freq Distill.`") achieves the best results across all metrics by providing a domain-compatible signal in the same feature space [5, 17, 51] as the backbone's representations, while explicitly counteracting the LF dominance with frequency-aware distillation from the final-layer teacher.

5.2. Analysis of Contrastive Learning Components

The detailed results for this analysis are available in Table 3. **Teacher selection.** When fixed to "`Negative: Random Layer`", using a "`Random Layer`" as the teacher underperforms, indicating that arbitrary references

do not consistently reflect the target representation to be learned. Moving the teacher to "Final Layer −1" or "Final Layer −2" further weakens either distortion or perceptual quality, aligning with the layer-depth-wise frequency progression: early layers stabilize LF, while the true "Final Layer" best captures converged HF details. Thus, the "Final Layer" teacher provides the most informative target for both IntraCL and InterCL. **Negative selection.** Fixing "Teacher: Final Layer", drawing the negative from the "Previous Layer" leads to weaker signals due to high correlation with the student. Sampling a "Random Layer" negative enforces comparisons across layer-depth, strengthens the margin, and—combined with the in-batch negatives in InterCL yields a more diverse and informative set of contrasts. The combination of "Final Layer" teacher and "Random Layer" negative achieves the strongest overall performance.

Table 3. **Ablation on contrastive components.**

Method	Quality Metrics		Perceptual Metrics	
	PSNR↑	LPIPS↓	MUSIQ↑	MANIQA↑
Part 1: Teacher Selection (Negative=Random Layer)				
Random Layer	22.63	0.39	67.23	0.469
Final Layer −1	22.15	0.40	67.45	0.476
Final Layer −2	22.62	0.39	65.82	0.438
Part 2: Negative Selection (Teacher=Final Layer)				
Previous Layer	22.23	0.39	66.44	0.445
Ours: Final Teacher + Random Neg.	**23.04**	**0.38**	**69.01**	**0.533**

5.3. Impact of Frequency-Specific Design

Table 4 compares the performance of different loss combinations for the LF and HF bands. **Individual frequency branches.** Applying only IntraCL on LF ("A") stabilizes globally shared structure but yields suboptimal improvement in HF details, giving limited perceptual gains. Applying only InterCL on HF ("D") sharpens instance-specific details and improves perceptual metrics, but the lack of LF stabilization can slightly hurt overall fidelity. Symmetrically, placing a single loss on the wrong band ("B" or "C") underutilizes the frequency-progressive hierarchy. **LF/HF combinations:** Pairing IntraCL for LF with InterCL for HF ("H") aligns distillation strength with the low-first, high-later hierarchy and yields the best distortion–perception trade-off. Reversing the assignment ("E") or using the same contrast on both bands ("F", "G") degrades either perceptual realism or fidelity, confirming that LF benefits from pairwise stabilization while HF benefits from instance discrimination with in-batch negatives.

5.4. Effectiveness of Adaptive Mechanisms

We validate the effectiveness of FAW and FAM in Table 5. **Adaptive weighting with FAW:** "FAW" computes normalized per-layer weights from student–teacher frequency discrepancies, aligning loss strength with the frequency-progressive hierarchy. Even without "FAM", "FAW" improves perceptual quality while maintaining distortion metrics. **Alignment gating with FAM:** "FAM" gates losses by the current LF/HF alignment degree to the final-layer teacher, reducing distillation strength when alignment is low (early layers) and increasing it as alignment improves (later layers). This stabilizes optimization and slightly improves both distortion and perception. **Combined effect:** Using both "FAW" and "FAM" yields the strongest overall performance, indicating that discrepancy-aware weighting and alignment-aware gating are complementary: "FAW" allocates distillation strength across frequencies and layer-depth, while "FAM" schedules its intensity according to layer-wise alignment. Together they best counter the LF bias and sharpen HF details without destabilizing early layers.

Table 4. **Ablation on the combination of contrastive losses for LF/HF bands.**

Variant	Loss Options		Quality Metrics		Perceptual Metrics	
	Low Freq.	High Freq.	PSNR↑	LPIPS↓	MUSIQ↑	MANIQA↑
Baseline	×	×	21.94	0.41	67.67	0.459
A	IntraCL	×	22.86	0.38	65.92	0.466
B	InterCL	×	22.62	0.41	66.22	0.479
C	×	IntraCL	22.27	0.39	67.98	0.509
D	×	InterCL	22.47	0.39	68.70	0.513
E	InterCL	IntraCL	22.21	0.41	67.56	0.484
F	InterCL	InterCL	21.90	0.41	67.21	0.465
G	IntraCL	IntraCL	22.85	0.39	65.25	0.435
H (Ours)	IntraCL	InterCL	**23.04**	**0.38**	**69.01**	**0.533**

Table 5. **Ablation study on the effectiveness of FAW and FAM.**

Variant	Adaptive Distillation		Quality Metrics		Perceptual Metrics	
	FAW	FAM	PSNR↑	LPIPS↓	MUSIQ↑	MANIQA↑
Baseline	×	×	21.94	0.41	67.67	0.459
CL only	×	×	23.04	0.38	69.01	0.533
FAW only	✓	×	22.93	0.38	70.50	0.552
FAM only	×	✓	23.06	0.37	69.43	0.540
FAW + FAM	✓	✓	**23.23**	**0.37**	**72.24**	**0.564**

6. Conclusion

FRAMER is a frequency-aware, layer-adaptive self-distillation framework for Real-ISR. It decomposes features with simple FFT masks and couples IntraCL and InterCL with FAW and FAM to deliver targeted optimization, prevent early-layer collapse, speed convergence, and enhance perceptual quality, without altering the backbone or inference. Across U-Net and DiT backbones and standard real-world benchmarks, FRAMER achieves consistent gains on both distortion and no-reference metrics, with ablations supporting the use of the final-layer feature map as teacher and random-layer plus in-batch negatives for HF.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. 2017. 2, 6

[2] 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. *Advances in Neural Information Processing Systems*, 37:55443–55469, 2024. 6, 7, 14

[3] Tariq Berrada, Pietro Astolfi, Melissa Hall, Marton Havasi, Yohann Benchetrit, Adriana Romero-Soriano, Karteek Alahari, Michal Drozdzal, and Jakob Verbeek. Boosting latent diffusion with perceptual objectives. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[4] 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. 2019. 6

[5] Wei-Chi Chen and Wei-Ta Chu. Sssd: Self-supervised self distillation. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2769–2776, 2023. 2, 3, 7

[6] Kun Cheng, Lei Yu, Zhijun Tu, Xiao He, Liyu Chen, Yong Guo, Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Hu. Effective diffusion transformer architecture for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2455–2463, 2025. 3

[7] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11472–11481, 2022. 2, 3, 6

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 3

[9] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 2

[10] Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy Ren, Chun-Le Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *ICCV 2025 Poster*, 2025. Exhibit Hall I #1755, Poster ID 534, Oct 22, 5:45–7:45 p.m. PDT. 1, 3, 6, 7, 14

[11] 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 *Forty-first international conference on machine learning*, 2024. 3, 6

[12] Fabian Falck, Teodora Pandeva, Kiarash Zahirnia, Rachel Lawrence, Richard Turner, Edward Meeds, Javier Zazo, and Sushrut Karmalkar. A fourier space perspective on diffusion models. *arXiv preprint arXiv:2505.11278*, 2025. 2, 3, 6

[13] Garas Gendy, Guanghui He, and Nabil Sabor. Diffusion models for image super-resolution: State-of-the-art and future directions. *Neurocomput.*, 617(C), 2025. 2

[14] Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. Div8k: Diverse 8k resolution image dataset. 2019. 6

[15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3, 7

[17] Jiho Jang, Seonhoon Kim, Kiyoon Yoo, Chaerin Kong, Jangho Kim, and Nojun Kwak. Self-distilled self-supervised representation learning. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2828–2838, 2023. 2, 3, 7

[18] Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. *arXiv preprint arXiv:2505.02831*, 2025. 3

[19] Thomas Jiralerspong, Berton Earnshaw, Jason Hartford, Yoshua Bengio, and Luca Scimeca. Shaping inductive bias in diffusion models through frequency-based noise control. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*, 2025. 3

[20] Tero Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019. 6

[21] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. 2021. 6

[22] Denis Kuznedelev, Valerii Startsev, Daniil Shlenskii, and Sergey Kastryulin. Does diffusion beat gan in image super resolution? *arXiv preprint arXiv:2405.17261*, 2024. 2

[23] Yueying Li, Hanbin Zhao, Jiaqing Zhou, Guozhi Xu, Tianlei Hu, Gang Chen, and Haobo Wang. FedSR: Frequency-aware enhancement for diffusion-based image super-resolution, 2025. 3

[24] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021. 7

[25] Leon Lin, Rodger Zhang, Jeya Maria Jose Valanarasu, Haoxiang Wang, Evangelos Gatti, Prajwal andpKalogerakis, and Vishal M Patel. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In *European Conference on Computer Vision (ECCV)*, 2024. 14

[26] 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: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LIX*, page 430–448, Berlin, Heidelberg, 2024. Springer-Verlag. 2

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3, 4, 11

[28] Xinyin Ma, Runpeng Yu, Songhua Liu, Gongfan Fang, and Xinchao Wang. Diffusion model is effectively its own teacher. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12901–12911, 2025. 3

[29] Tejaswini Medi, Hsien-Yi Wang, Arianna Rampini, and Margret Keuper. Missing fine details in images: Last seen in high frequencies. *arXiv e-prints*, pages arXiv–2509, 2025. 2

[30] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 6

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2, 3, 4, 6

[32] 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*, pages 10684–10695, 2022. 2, 3, 6

[33] 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*, pages 10684–10695, 2022. 14

[34] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. arxiv 2014. *arXiv preprint arXiv:1412.6550*, 2014. 3

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 4, 6

[36] 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, 2022. 2, 3

[37] Shoaib Meraj Sami, Md Mahedi Hasan, Jeremy Dawson, and Nasser Nasrabadi. Hf-diff: High-frequency perceptual loss and distribution matching for one-step diffusion-based image super-resolution. *arXiv preprint arXiv:2411.13548*, 2024. 3

[38] Zhuoyi Shen, Maoyu Mao, and Pengfei Fan. A primary comparison of diffusion models and generative adversarial networks for image synthesis. In *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI)*, page 225–234, New York, NY, USA, 2024. Association for Computing Machinery. 2

[39] 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 Computer Vision and Pattern Recognition Conference*, pages 2333–2343, 2025. 1, 3, 7, 14

[40] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 3, 7

[41] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. 2017. 6

[42] Yuhao Wan, Peng-Tao Jiang, Qibin Hou, Hao Zhang, Jinwei Chen, Ming-Ming Cheng, and Bo Li. Controlsr: Taming diffusion models for consistent real-world image super resolution. *arXiv preprint arXiv:2410.14279*, 2024. 2

[43] 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*, pages 1905–1914, 2021. 2, 3, 4, 6, 11, 12

[44] Xingjian Wang, Li Chai, and Jiming Chen. Frequency-domain refinement with multiscale diffusion for super resolution. *arXiv preprint arXiv:2405.10014*, 2024. 3

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. 2004. 6

[46] 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. 2020. 6

[47] Damion Woods and Peter Bloem. Self-distillation for diffusion models, 2024. 3

[48] 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*, pages 25456–25467, 2024. 2, 6, 7, 14

[49] 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. 2022. 6

[50] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in Neural Information Processing Systems*, 36:13294–13307, 2023. 7

[51] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3713–3722, 2019. 2, 7

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. 2018. 2, 6

[53] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11953–11962, 2022. 2, 3, 7

FRAMER: Frequency-Aligned Self-Distillation with Adaptive Modulation Leveraging Diffusion Priors for Real-World Image Super-Resolution

Supplementary Material

A. Implementation and Algorithm Details

A.1. Training Algorithm

Algorithm 1 FRAMER Training Scheme

Require: HR images R, Diffusion Model ϵ_θ, Teacher Layer n, Predefined Frequency Masks $M_{\text{LF}}, M_{\text{HF}}$ (radius $r = 0.2\%$)

Data Preparation:
  Generate $I_{\text{LR}} \leftarrow \text{Degradation}(R)$ {See Sec. A.2}
  Generate Caption $C \leftarrow \text{LLaVA}(I_{\text{LR}})$
  Sample timestep $t \sim [1, T]$, noise $Z \sim \mathcal{N}(0, \mathbf{I})$
Forward Pass:
  Encode HR image R to latent z_0; Get noisy latent Z_t using Z, t
  Feed $(Z_t, t, I_{\text{LR}}, C)$ to ϵ_θ
  Extract Teacher Feature $\mathbf{F}^{(n)}$ and Student Features $\{\mathbf{F}^{(i)}\}_{i=1}^{N}$
for each layer i in Students **do**
  Frequency Decomposition:
  $\mathbf{F}_{\text{LF}}^{(i)}, \mathbf{F}_{\text{LF}}^{(n)} \leftarrow \text{FFT}(\mathbf{F}^{(i)}, \mathbf{F}^{(n)}) \odot M_{\text{LF}}$
  $\mathbf{F}_{\text{HF}}^{(i)}, \mathbf{F}_{\text{HF}}^{(n)} \leftarrow \text{FFT}(\mathbf{F}^{(i)}, \mathbf{F}^{(n)}) \odot M_{\text{HF}}$
  Compute Losses & Modulation:
  Calculate $\mathcal{L}_{\text{IntraCL}}^{(i)}$ using Eq. 1 {Stabilize shared structure}
  Calculate $\mathcal{L}_{\text{InterCL}}^{(i)}$ using Eq. 2 {Sharpen instance details}
  Compute weights $\mathbf{w}^{(i)}$ via **FAW** (Eq. 5)
  Compute alignment $a^{(i)}$ via **FAM** (Eq. 7)
  $\mathcal{L}_{\text{FRAMER}}^{(i)} \leftarrow$ Weighted Sum (Eq. 9)
end for
Optimization:
$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{noise}} + \sum_i \mathcal{L}_{\text{FRAMER}}^{(i)}$
Update θ via Backpropagation on $\mathcal{L}_{\text{total}}$

Algorithm 2 FRAMER Inference Scheme

Require: LR Image I_{LR}, Pre-trained Diffusion Model ϵ_θ

Preparation:
  Generate Caption $C \leftarrow \text{LLaVA}(I_{\text{LR}})$
  Sample noise $Z_T \sim \mathcal{N}(0, \mathbf{I})$ {Initialize at **Target HR Latent Size**}
Reverse Sampling Process:
for $t = T, \ldots, 1$ **do**
  $\epsilon_t \leftarrow \epsilon_\theta(Z_t, t, I_{\text{LR}}, C)$ {Predict noise conditioned on LR}
  $Z_{t-1} \leftarrow \text{Sampler}(Z_t, \epsilon_t)$ {Denoise towards HR latent}
end for
Reconstruction:
return HR Image $R \leftarrow \text{Decode}(Z_0)$ {Maps latent to **HR Pixel Space**}

We outline the detailed training procedure of FRAMER in **Algorithm 1**. As described in the main paper, FRAMER is designed as a plug-and-play training strategy that leverages diffusion priors without altering inference.

Training Phase. We first synthesize Low-Resolution (LR) inputs I_{LR} from High-Resolution (HR) images R using the Real-ESRGAN degradation pipeline [43]. Concurrently, we utilize LLaVA [27] to generate a descriptive caption for each I_{LR}. The diffusion model takes $(I_{\text{LR}}, Z_T, \text{caption})$ as input. During the forward process, we extract feature maps from intermediate layers (students) and the final-layer feature map (teacher), which serves as the target representation.

As detailed in Sec. 3.1 of the main paper, we decompose these features into Low-Frequency (LF) and High-Frequency (HF) bands via 2D FFT using binary masks. We then compute the auxiliary distillation losses:

- **IntraCL (Sec. 3.2):** Applied to the LF band to stabilize globally shared structures. It compares a student only against its teacher and a random-layer negative within the same network, avoiding false negatives common in batch-based contrastive learning.
- **InterCL (Sec. 3.3):** Applied to the HF band to sharpen instance-specific details. It uses in-batch negatives and random-layer negatives to promote instance discrimination and layer-wise progression.

These objectives are modulated by **FAW** (Sec. 3.4), which weights distillation based on the relative frequency difference to the final layer, and gated by **FAM** (Sec. 3.5), which controls distillation strength according to the student-

Table 6. **Hyperparameters for the Degradation Pipeline.**

Degradation Type	Parameter Settings
First Degradation Stage	
Blur Kernel Size	21×21
Blur Sigma	$[0.2, 3.0]$
Blur Kernel Types	iso, aniso, generalized_iso, generalized_aniso, plateau_iso, plateau_aniso
Sinc Probability	0.1
Resize Range	$[0.15, 1.5]$ (Up/Down/Keep)
Gaussian Noise	Prob: 0.5, Sigma: $[1, 30]$
Poisson Noise	Scale: $[0.05, 3.0]$
JPEG Compression	Quality: $[30, 95]$
Second Degradation Stage	
Blur Kernel Size	11×11
Blur Sigma	$[0.2, 1.5]$
Sinc Probability	0.1
Resize Range	$[0.3, 1.2]$ (Up/Down/Keep)
Gaussian Noise	Prob: 0.5, Sigma: $[1, 25]$
Poisson Noise	Scale: $[0.05, 2.5]$
JPEG Compression	Quality: $[30, 95]$
Final Processing	
Final Sinc Prob	0.8
Crop Size	512

teacher alignment. The total objective combines the standard noise-prediction loss with these frequency-aligned distillation terms (Eq. 10).

Inference Phase. We summarize the inference procedure in **Algorithm 2**. During inference, FRAMER introduces **no computational overhead**. All auxiliary heads and loss computations are strictly training-only and removed at test time. We simply generate a caption for the input LR image using LLaVA and perform standard sampling with the optimized diffusion backbone.

A.2. Degradation Pipeline Details

We follow the high-order degradation process used in Real-ESRGAN [43] to synthesize training pairs. The specific parameters used in our implementation are summarized in Table 6.

A.3. Computational Cost Analysis

To evaluate the computational overhead introduced by FRAMER, we compare the training memory usage and time per iteration against the baseline DiT4SR model. All measurements were conducted on a single NVIDIA H200 GPU with a batch size of 16.

As shown in Fig. 6, FRAMER incurs a marginal increase in computational resources due to the additional FFT decomposition and auxiliary loss calculations (IntraCL, InterCL). Specifically:

- **Memory Usage:** Increases by approximately **3.0%** (from 87.03 GB to 89.65 GB).
- **Training Time:** Increases by approximately **6.9%** per iteration (from 1.01s to 1.08s).

Despite these slight increases during training, we consider this cost negligible given the significant improvements in convergence stability and final quality. **Crucially, FRAMER introduces zero overhead during inference.** Since the auxiliary heads and frequency-aware losses are strictly removed after training, the inference speed and memory consumption remain identical to the original backbone.

B. Analysis of Training Dynamics and Hierarchy

B.1. Validation of "Low-first, High-later" Hierarchy

To empirically validate the depth-wise "low-first, high-later" hierarchy discussed in Sec. 1 and Sec. 3.1 of the main paper, we compare the layer-wise feature alignment of our FRAMER-trained model against the baseline DiT4SR model. Fig. 7 plots the cosine similarity between intermediate layer features and the final-layer teacher features for both LF and HF components at two distinct noise timesteps ($t = 300$ and $t = 700$).

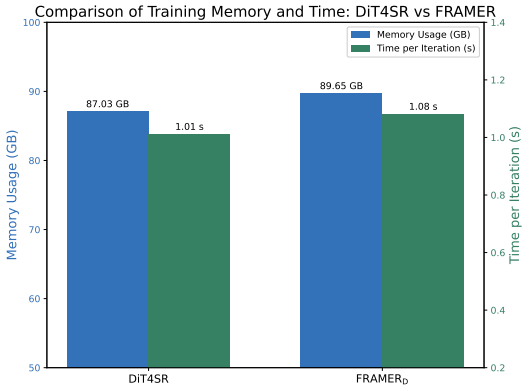


Figure 6. **Comparison of Training Cost (Memory and Time).** We measure the GPU memory usage and time per iteration for DiT4SR and FRAMER$_D$ on an **NVIDIA H200 GPU with a batch size of 16**. FRAMER introduces only a marginal training overhead (3% memory, 7% time) while maintaining identical inference costs due to its plug-and-play nature.

LF Stability (Blue Lines). As shown by the blue curves, both models achieve relatively high alignment for LF components across all layers. However, FRAMER (dashed blue) consistently maintains higher similarity scores than the baseline (solid blue), particularly in the earlier layers. This indicates that our **IntraCL** successfully stabilizes the global structure early in the network depth, preventing structural distortions during the denoising process.

HF Acceleration (Red Lines). The most significant difference is observed in the HF components. The baseline DiT4SR (solid red) exhibits a distinct "low-first, high-later" behavior: HF similarity remains near zero for the majority of the network depth and only spikes abruptly in the final few layers. This confirms our observation in the main paper that the standard noise-prediction loss leaves intermediate layers under-optimized for fine details. In contrast, FRAMER (dashed red) demonstrates a much earlier rise in HF alignment, starting to increase significantly around Layer 10. This proves that our **InterCL**, modulated by FAW and FAM, effectively "pre-aligns" intermediate layers to the high-frequency details of the teacher. By mitigating the depth-wise delay in HF learning, FRAMER enables the model to dedicate more capacity to refining textures and edges throughout the network.

B.2. Visual Analysis of Training Stability

To verify the effectiveness of FAW and FAM in stabilizing the optimization process and preventing potential model collapse, we conduct a visual analysis of the training dynamics during the **initial training phase** (1k–5k iterations). Fig. 8 compares the reconstruction quality across different configurations.

Prevention of Early-Stage Collapse. In the very early

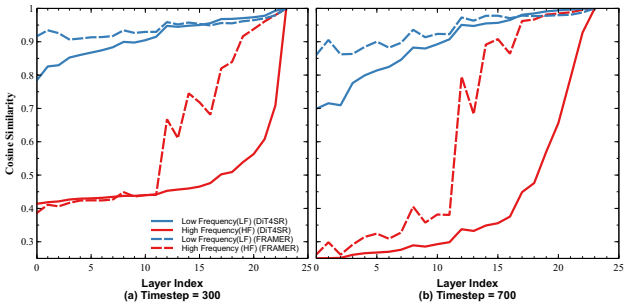


Figure 7. **Layer-wise cosine similarity comparison between the baseline (DiT4SR) and FRAMER.** We measure the similarity of intermediate features to the final-layer teacher features for LF (blue) and HF (red) bands. (a) At $t = 300$ and (b) $t = 700$, the baseline (solid lines) shows a delayed response for HF components, validating the "low-first, high-later" hierarchy described in the main paper. In contrast, FRAMER (dashed lines) significantly accelerates HF alignment in intermediate layers (Layer 10–20), demonstrating that our frequency-aligned distillation effectively counteracts the spectral bias.

stages (1k–2k iterations), the "Only Distill" model and single-module variants often exhibit signs of training instability, producing chaotic artifacts or failing to form coherent structures. This suggests that aggressive self-distillation without adaptive modulation can lead to optimization difficulties or early-stage collapse, as the student is forced to mimic the teacher before establishing basic features.

Consistent and Stable Optimization. In contrast, **FRAMER (Distill + FAW, FAM)** demonstrates a stable and consistent progression throughout these initial steps. Even at 5k iterations, the model establishes a solid structural foundation with significantly reduced noise compared to other settings. This visual evidence confirms that FAM effectively gates large, unstable gradients when student-teacher alignment is low, while FAW ensures balanced frequency optimization, collectively securing a stable training trajectory from the start.

Figure 8. **Visual analysis of training stability during the initial phase.** We compare the reconstruction quality from 1k to 5k iterations. While the baseline and single-module variants show signs of instability or incoherent structures, our full method (**Distill + FAW, FAM**) demonstrates a stable optimization trajectory, effectively preventing early-stage model collapse. Red arrows indicate artifacts within each generated image. *Best viewed in Zoom.*

C. Additional Qualitative Results

We provide comprehensive visual comparisons to further demonstrate the effectiveness of FRAMER. We categorize the evaluation into two groups: (1) datasets with Ground Truth (GT) references (RealSR, DrealSR) and (2) datasets without Ground Truth (RealLR200, RealLQ250), representing real-world "in-the-wild" scenarios.

C.1. Comparisons on Datasets with Ground Truth

Fig. 11 presents comparisons on the RealLR200 and RealLQ250 datasets, which consist of low-quality real-world images with unknown degradations and no ground truth. These scenarios are particularly challenging due to severe artifacts and the risk of hallucination. Comparison methods often fail to remove heavy noise or generate unnatural artifacts (e.g., distorted facial features or blurred textures). **FRAMER** demonstrates robust generalization capabilities. For instance, in the vintage portraits, FRAMER effectively suppresses noise while enhancing facial details (eyes, hair strands) without creating uncanny artifacts. Similarly, in the animal images (parrots, frog), it restores the intricate textures of feathers and skin that are often lost by other methods. This highlights FRAMER's ability to generate perceptually pleasing and natural results even in the absence of ground truth guidance.

D. User Study

To complement the quantitative metrics and verify the subjective superiority of our method, we conducted a user preference study.

Experimental Setup. We invited 15 participants to evaluate the visual quality of the restored images. The study comprised a total of 30 distinct scenes randomly selected from the RealSR and DrealSR, RealLR200, RealLQ250 datasets. Participants were asked to select the best image among the comparison methods based on three criteria: (1) *Fidelity* (faithfulness to ground truth details and structure), (2) *Perceptual Quality* (sharpness, naturalness, and lack of artifacts), and (3) *Overall Quality* (general preference considering both fidelity and realism).

Architecture-wise Comparison. To ensure a rigorous and fair evaluation, we divided the study into two dis-

Table 7. **User Study Results (Win Rate %).** The values indicate the percentage of votes where each method was selected as the best. We evaluate the win rate within each architecture group (U-Net and DiT) to ensure a fair comparison. **Bold** indicates the best performance.

	U-Net-based Methods			DiT-based Methods		
Metrics	SeeSR [48]	PiSA-SR [39]	**FRAMER$_U$**	DreamClear [2]	DiT4SR [10]	**FRAMER$_D$**
Fidelity (↑)	23.3	20.7	**56.0**	13.3	33.3	**53.3**
Perceptual Quality (↑)	12.0	16.0	**72.0**	13.3	13.3	**73.3**
Overall Quality (↑)	13.0	18.2	**68.8**	6.7	26.7	**66.7**

tinct tracks based on the backbone architecture: (1) **U-Net-based comparison** (SeeSR, PiSA-SR, vs. FRAMER$_U$) and (2) **DiT-based comparison** (DreamClear, DiT4SR, vs. FRAMER$_D$). Since different backbone architectures possess different baseline capabilities, this grouped comparison is crucial to isolate the performance gains contributed purely by our FRAMER training framework, rather than the architectural differences.

Results and Analysis. The results of the user study are summarized in Table 7. The values represent the percentage of votes where each method was selected as the best.

- In the **U-Net group**, FRAMER$_U$ significantly outperforms the baselines, securing **56.0%** of votes in Fidelity and **72.0%** in Perceptual Quality. This indicates that users clearly distinguish the enhanced detail and structural stability provided by our method.
- In the **DiT group**, the preference for FRAMER$_D$ is even more pronounced, reaching **73.3%** in Perceptual Quality. This suggests that our frequency-aligned distillation effectively unlocks the potential of the DiT backbone for generating high-frequency details that are visually pleasing to human observers.

Overall, FRAMER consistently achieves the highest preference rates across all metrics and architectures, confirming that our method produces results that are not only quantitatively superior but also perceptually more realistic and faithful.

E. Limitations and Future Work

While FRAMER demonstrates state-of-the-art performance in restoring high-frequency details and maintaining perceptual quality, it is not entirely exempt from the inherent challenges of generative diffusion models.

Generative Artifacts. As noted in prior studies [33], generative models tend to hallucinate semantic details or textures that do not exist in the original scene, especially when the input low-resolution image suffers from extreme degradation. Although FRAMER significantly mitigates this issue compared to pure generative baselines by enforcing feature consistency via self-distillation, minor artifacts or unfaithful texture synthesis may still occur. For instance, as shown in Fig. 9, while our method reconstructs the rope texture with high definition compared to the blurry or artifact-prone baselines, the specific twisting pattern may slightly diverge from the exact pixel-structure of the Ground Truth.

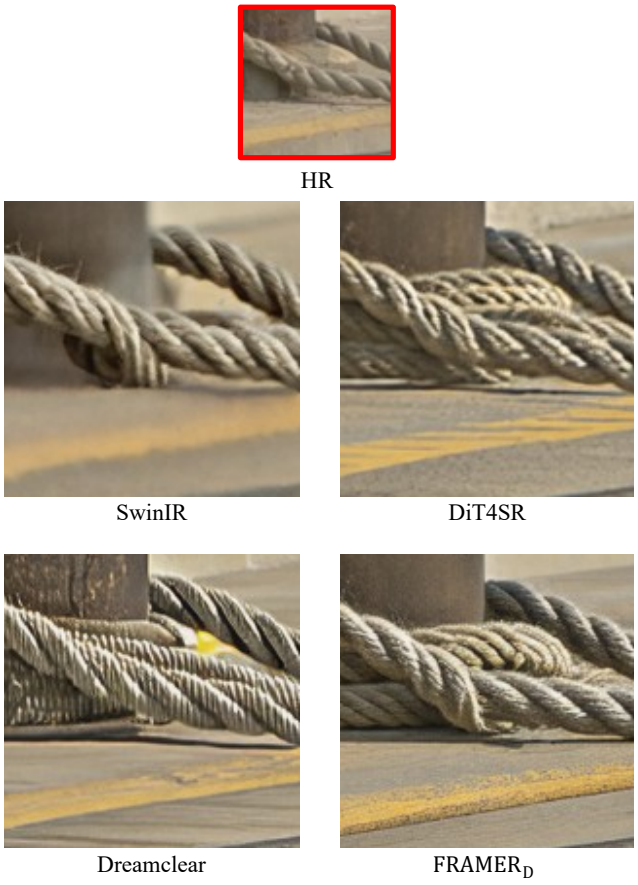


Figure 9. **Visual illustration of fidelity limitations.** We compare the restoration of challenging rope textures. While FRAMER$_D$ produces results that are perceptually far superior and sharper than baselines (SwinIR, DiT4SR, DreamClear), the generated fine details may exhibit slight structural deviations from the Ground Truth (HR). This illustrates the inherent trade-off between perceptual realism and pixel-wise fidelity in generative super-resolution.

Future Directions. To further address the stochastic nature of diffusion-based upscaling and minimize hallucinations, future work could explore integrating frequency-constrained sampling strategies. For instance, adapting training-free inference techniques like **FouriScale** [25], which manipulates frequency components during the reverse sampling process to ensure structural rigidity, could complement our training-time frequency alignment. Combining FRAMER's robust representation learning with such inference-time constraints represents a promising avenue for achieving hallucination-free, high-fidelity super-resolution.

Figure 10. **Qualitative comparisons on datasets with Ground Truth (RealSR, DrealSR).** We compare FRAMER against state-of-the-art methods (SwinIR, ResShift, SeeSR, PiSA-SR, DreamClear, DiT4SR). We highlight specific failure cases in baseline methods: **Red arrows** indicate structural errors (e.g., hallucinations, object distortion), while **Yellow arrows** point to textural defects (e.g., over-sharpening, blur, noise). In contrast, our methods (FRAMER$_U$, FRAMER$_D$) consistently mitigate these artifacts, producing sharper edges and faithful textures closely aligning with the HR references.

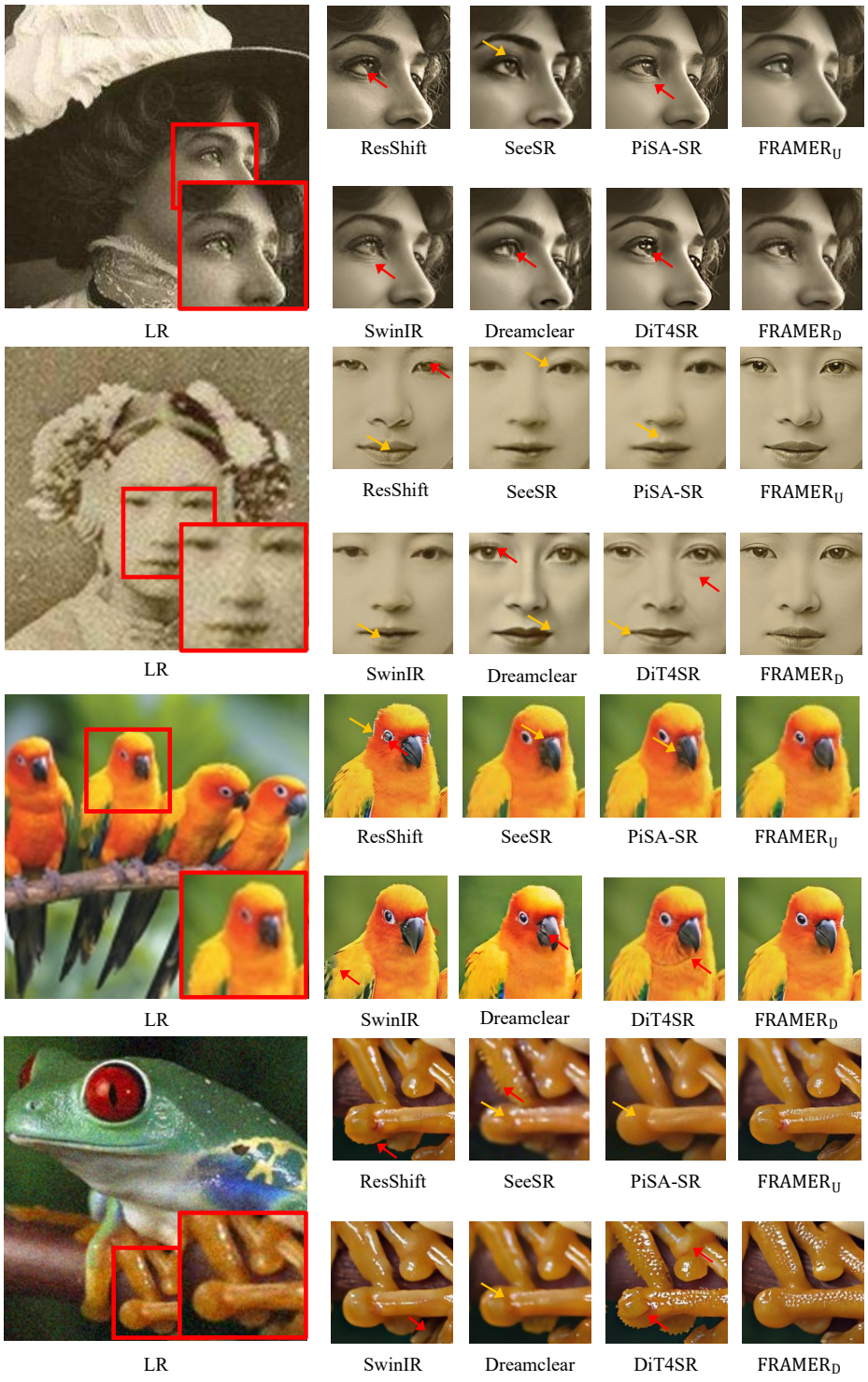


Figure 11. **Qualitative comparisons on datasets without Ground Truth (RealLR200, RealLQ250).** In these real-world scenarios with unknown degradations, baseline methods often suffer from severe degradations marked by arrows: **Red** indicates structural failures (e.g., hallucinations, object crushing), and **Yellow** indicates textural anomalies (e.g., over-sharpening, residual noise). FRAMER demonstrates superior perceptual quality by effectively balancing noise suppression with detail generation, avoiding these common pitfalls observed in competing methods.