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 $t = 700$).

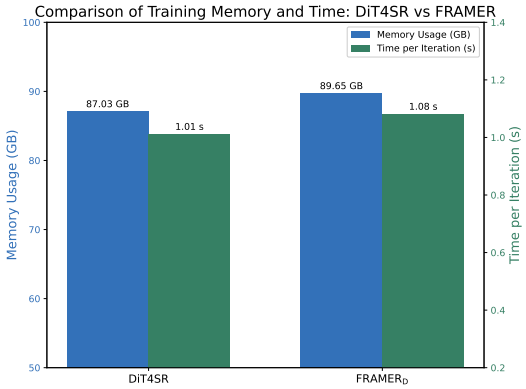


Figure 6. **Comparison of Training Cost (Memory and Time).** We measure the GPU memory usage and time per iteration for DiT4SR and FRAMER$_D$ on an **NVIDIA H200 GPU with a batch size of 16**. FRAMER introduces only a marginal training overhead (3% memory, 7% time) while maintaining identical inference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

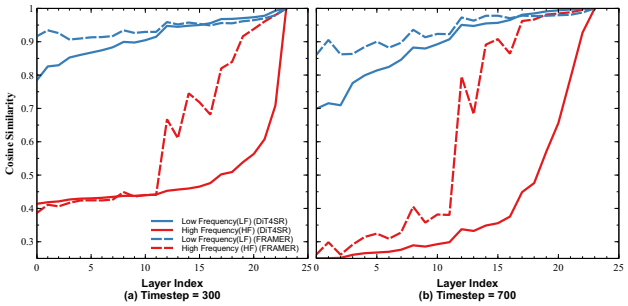


Figure 7. **Layer-wise cosine similarity comparison between the baseline (DiT4SR) and FRAMER.** We measure the similarity of intermediate features to the final-layer teacher features for LF (blue) and HF (red) bands. (a) At $t = 300$ and (b) $t = 700$, the baseline (solid lines) shows a delayed response 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.

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.*

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.

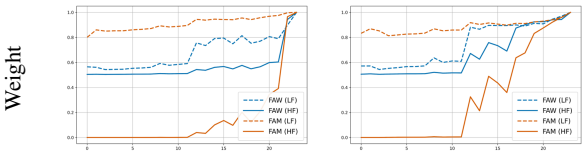
C. Additional Analysis on Adaptive Modulation (FAW and FAM)

In the main paper, we introduced the Frequency-based Adaptive Weight (FAW) and Frequency-based Alignment Modulation (FAM) to dynamically control the self-distillation process. In this section, we provide further empirical evidence and visualizations to support their design and complementary roles.

C.1. FAW and FAM Weight Visualizations

To demonstrate that our adaptive modules operate as intended according to the depth-wise frequency hierarchy, we visualize the layer-wise weights during different training phases in Fig. 9.

As shown in the Fig. 9, HF supervision is weak across most layers during the early training phase. However, as training progresses to the late phase, the HF supervision strengthens across more layers. This behavior perfectly aligns with our motivation: the network first prioritizes stabilizing the globally shared LF structures and subsequently focuses on refining the instance-specific HF details.



(a) Early training phase (b) Late training phase

Figure 9. **Visualization of FAW/FAM weights across layers.** (a) Early training phase and (b) Late training phase. The visualizations confirm the intended behavior: HF supervision is relatively weak across most layers early on, and strengthens across more layers later in the training process.

C.2. Complementary Roles of FAW and FAM

We also noted the synergistic relationship between FAW and FAM. Fig. 10 provide qualitative and quantitative ablation comparisons to support this claim.

Using FAW alone rigidly scales the loss without considering the student-teacher alignment, which often leads

to oversharpening artifacts. Conversely, using FAM alone lacks the frequency-specific dynamic weighting, resulting in insufficient perceptual quality. By combining both, FRAMER yields more coherent structures with balanced perceptual realism. This trend is consistently reflected in the perceptual metrics (NIQE, MANIQA, and MUSIQ), where the combined use of FAW and FAM strictly achieves the best scores.

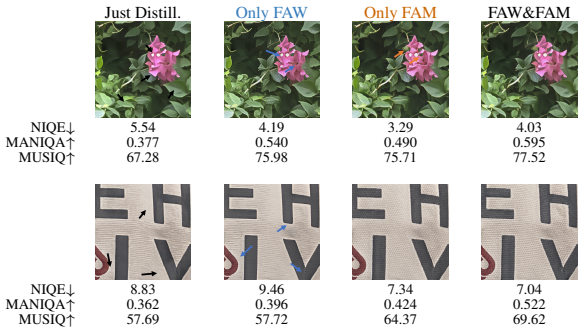


Figure 10. **Qualitative comparison of adaptive modules.** Using FAW alone often causes oversharpening artifacts, while using FAM alone results in insufficient perceptual quality. FAW and FAM together produce the most coherent structures and balanced perceptual realism.

D. Plug-and-Play Generalization to One-step and GAN-augmented Diffusion Models

In the main paper, we primarily demonstrated the effectiveness of FRAMER on multi-step diffusion backbones (U-Net and DiT). To further validate the versatility and generalization capability of our proposed framework, we extend the evaluation of FRAMER to both one-step and GAN-augmented diffusion models.

Specifically, we apply FRAMER to a representative one-step method, **TSD-SR**, and a GAN-augmented method, **SupResDiffGA**, under strictly matched training protocols. As shown in Table 7, FRAMER generalizes exceptionally well across these different model architectures and training paradigms.

On the RealSR dataset, applying FRAMER to the GAN-augmented baseline (SupResDiffGA) achieves state-of-the-art fidelity metrics (PSNR, SSIM, LPIPS), while DiT-based FRAMER models attain state-of-the-art perceptual quality. Furthermore, applying FRAMER to the one-step baseline (TSD-SR) consistently boosts both fidelity and perceptual metrics (NIQE, MANIQA, MUSIQ). Similar consistent improvements are observed on the RealLQ250 dataset. This confirms that the internal LF bias and the depth-wise frequency hierarchy are broadly shared optimization challenges in diffusion-based restoration, and FRAMER serves as an effective, backbone-agnostic solution.

Table 7. **Quantitative comparison on one-step and GAN-augmented diffusion models.** FRAMER consistently improves both fidelity and perceptual metrics across different diffusion paradigms. The green percentages indicate the relative improvement over the respective baselines. (Note: Fidelity metrics are omitted for RealLQ250 as no ground-truth is available).

Dataset	Model	PSNR↑	SSIM↑	LPIPS↓	NIQE↓	MANIQA↑	MUSIQ↑
RealSR	TSD-SR	23.45	0.696	0.383	5.163	0.507	71.08
	FRAMER$_{\text{TSD-SR}}$	**24.46** (4.3%)	**0.732** (5.1%)	**0.345** (9.8%)	**4.295** (16.8%)	**0.595** (17.3%)	**73.63** (3.6%)
	SupResDiffGA	24.36	0.697	0.472	7.294	0.198	39.57
	FRAMER$_{\text{SupResDiffGA}}$	**25.72** (5.6%)	**0.751** (7.8%)	**0.416** (11.8%)	**6.229** (14.6%)	**0.230** (15.9%)	**41.59** (5.1%)
RealLQ250	TSD-SR				7.418	0.589	69.63
	FRAMER$_{\text{TSD-SR}}$				**6.988** (5.8%)	**0.624** (6.0%)	**71.38** (2.5%)
	SupResDiffGA	*No ground-truth available*			5.071	0.343	57.51
	FRAMER$_{\text{SupResDiffGA}}$				**4.624** (8.8%)	**0.375** (9.2%)	**62.98** (9.5%)

E. 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.

E.1. Comparisons on Datasets with Ground Truth

Fig. 13 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.

F. 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).

Table 8. **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**

Architecture-wise Comparison. To ensure a rigorous and fair evaluation, we divided the study into two distinct 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 8. 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.

G. 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. 11, 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.

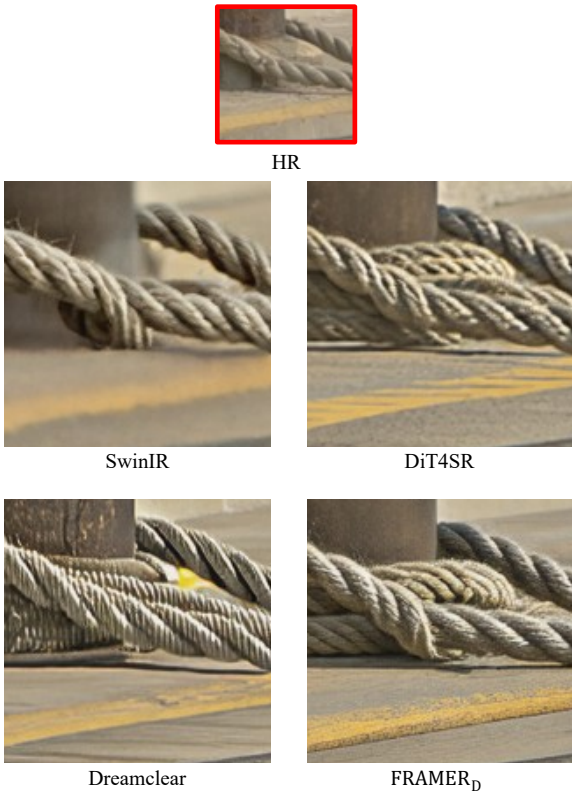


Figure 11. **Visual illustration of fidelity limitations.** We compare the restoration of challenging rope textures. While FRAMER$_D$ produces results that are perceptually far superior and sharper than baselines (SwinIR, DiT4SR, DreamClear), the generated fine details may exhibit slight structural deviations 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 12. **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.

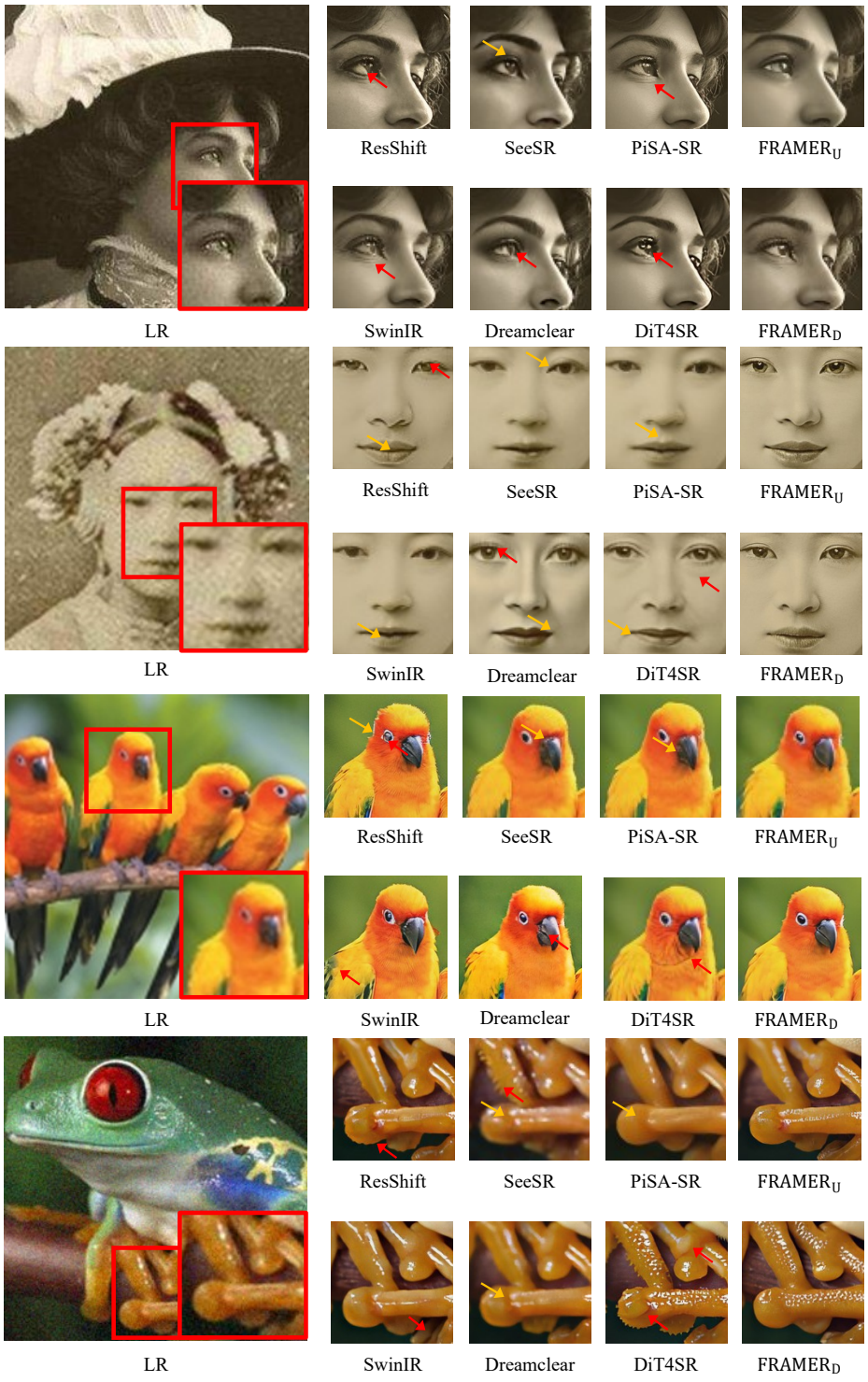


Figure 13. **Qualitative comparisons on datasets without Ground Truth (RealLR200, RealLQ250).** In these real-world scenarios with unknown degradations, baseline methods often suffer from severe degradations marked by arrows: **Red** indicates structural failures (e.g., hallucinations, object crushing), 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.