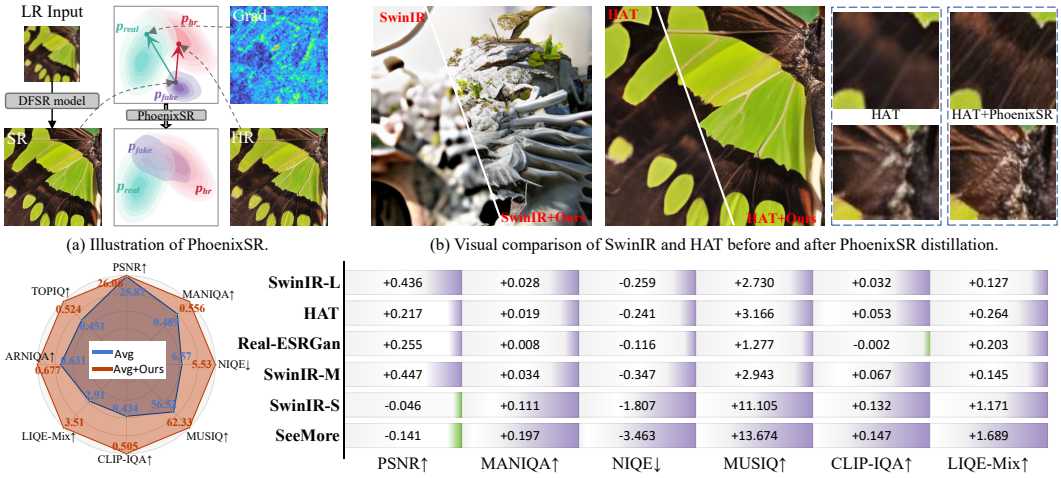
PHOENIXSR: GENERATIVE HETEROGENEOUS DISTILLATION UNLEASHES EFFICIENT MODELS FOR REAL-WORLD SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review



(a) Illustration of PhoenixSR.

(b) Visual comparison of SwinIR and HAT before and after PhoenixSR distillation.

(c) Per-model and average performance improvements.

Figure 1: **Overview and performance of PhoenixSR.** (a) PhoenixSR transfers diffusion generative priors to diffusion-free SR students through generative heterogeneous distillation. (b) Visual comparisons of SwinIR and HAT before and after PhoenixSR distillation. (c) Average and model-wise performance improvements brought by PhoenixSR on RealSR.

ABSTRACT

Real-world image super-resolution (SR) requires recovering perceptually realistic high-resolution images from complex low-resolution observations while preserving faithful content. Diffusion-based SR benefits from strong generative priors but incurs substantial computational overhead, whereas feed-forward CNN and Transformer SR models are efficient yet often struggle to recover realistic high-frequency details. This motivates a natural question: *can diffusion priors be transferred to existing diffusion-free SR networks without introducing diffusion components at inference time?* To this end, we propose **PhoenixSR**, a generative heterogeneous distillation framework that transfers diffusion priors to independently designed feed-forward SR networks through score-based distribution matching. Rather than aligning heterogeneous features or imitating sampled diffusion outputs, PhoenixSR uses the pretrained diffusion model as distribution-level supervision, while paired SR supervision preserves reconstruction fidelity. To make distribution matching effective for fidelity-sensitive SR, we introduce **Heterogeneous Distribution Adaptation**, which adapts the target score to the SR domain, improves tracking of the evolving student distribution, and anchors training with paired supervision. We further employ **Directional Reliability Weighting**, a lightweight residual-consistency-based reweighting strategy that reduces unstable distributional guidance. All diffusion-related components are removed after training, leaving the original student architecture and inference cost unchanged. Experiments on three SR benchmarks and six feed-forward backbones, including SwinIR, HAT, Real-ESRGAN, and SeeMoRe, show consistent perceptual improvements with largely preserved reconstruction fidelity.

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

1 INTRODUCTION

Single image super-resolution (SISR) aims to reconstruct a high-resolution (HR) image from its low-resolution (LR) observation and has been extensively studied in computer vision (Dong et al., 2014; 2016a; Kim et al., 2016; Lim et al., 2017; Zhang et al., 2018b; Dai et al., 2019; Zhang et al., 2019; Niu et al., 2020; Liang et al., 2021; Chen et al., 2023). While early methods mainly consider synthetic degradations, real-world images often suffer from complex and unknown degradation processes, including blur, sensor noise, compression, and camera processing. Real-world SR is therefore particularly challenging, as the restored image should be perceptually realistic while remaining faithful to the input content (Cai et al., 2019; Zhang et al., 2021; Wang et al., 2021).

Recent diffusion-based super-resolution (DBSR) methods achieve impressive perceptual quality by exploiting generative priors learned from large-scale natural image distributions (Saharia et al., 2023; Rombach et al., 2022; Wang et al., 2024a; Yue et al., 2023; Lin et al., 2024; Wu et al., 2024b; Yang et al., 2024; Yu et al., 2024). However, iterative diffusion sampling requires repeated network evaluations and introduces substantial computation and latency (Ho et al., 2020; Song et al., 2021; Nichol & Dhariwal, 2021; Lu et al., 2022). Although recent methods reduce the number of sampling steps (Wang et al., 2024b; Wu et al., 2024a; Chen et al., 2025), their inference models still rely on diffusion backbones or diffusion-derived generation pipelines. In contrast, diffusion-free SR (DFSR) models directly reconstruct HR images through a single feed-forward pass (Dong et al., 2014; Kim et al., 2016; Lim et al., 2017; Zhang et al., 2018b; Liang et al., 2021; Chen et al., 2023; Kong et al., 2022; Li et al., 2020; Sun et al., 2023). They are substantially more efficient, but often lack strong generative priors and tend to produce over-smoothed details under severe degradations. This raises a fundamental question: *can the generative prior of a pretrained diffusion model be transferred to existing diffusion-free SR networks without introducing diffusion components at inference time?*

Knowledge distillation provides a natural way to transfer such knowledge (Hinton et al., 2015; Romero et al., 2015; Gou et al., 2021). However, conventional distillation is not well suited to diffusion-to-DFSR transfer because the teacher and student follow very different generation processes. Feature-level distillation usually relies on meaningful correspondence between intermediate representations (Romero et al., 2015; Heo et al., 2019), whereas diffusion features depend on noisy states and denoising dynamics, while DFSR features are optimized for deterministic reconstruction. As shown in Fig. 2, DBSR and DFSR exhibit limited feature correspondence. Output distillation also has an inherent limitation: diffusion-based restoration may generate perceptually plausible details that do not exactly agree with the paired HR target, and directly imitating them may transfer stochastic or teacher-specific textures. Recent work such as DTKD (Park & Song, 2026) transfers diffusion knowledge to feed-forward students through output supervision. In contrast, we explore whether the diffusion model can act as a *distribution-level teacher*, avoiding both feature alignment and imitation of sampled SR outputs.

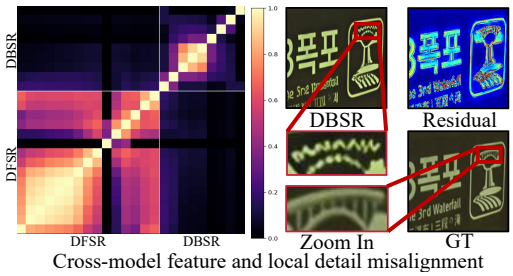
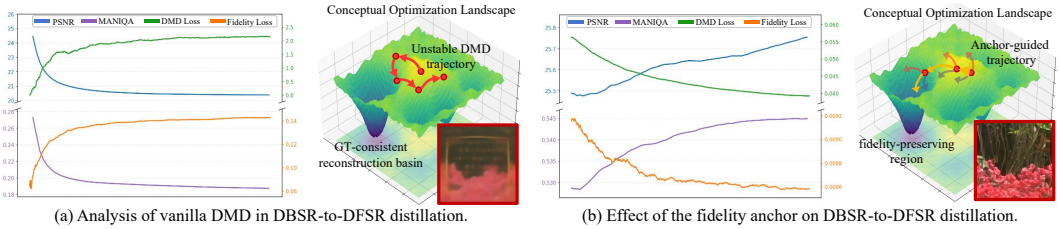


Figure 2: **Motivation.** CKA (Kornblith et al., 2019) (left) reveals weak feature correspondence between DBSR and DFSR, while residual and zoom-in comparisons (right) show local discrepancies between DBSR outputs and paired HR targets.

Distribution Matching Distillation (DMD) (Yin et al., 2024b;a) is particularly suitable for this purpose because real–fake score differences can supervise a differentiable generator without requiring architectural correspondence. This makes it possible to use a pretrained diffusion model to train independently designed DFSR networks such as SwinIR, HAT, and SeeMoRe. However, directly applying DMD to SR is insufficient. SR is a paired restoration task in which each LR input should correspond to a specific HR target, whereas distribution matching mainly constrains the overall output distribution. As shown in Fig. 3(a), vanilla DMD causes unstable optimization and rapid PSNR degradation. Paired reconstruction supervision substantially stabilizes training and prevents the diffusion prior from overriding input-specific content (Fig. 3(b)). Thus, distribution-level diffusion distillation requires explicit adaptation to the fidelity-sensitive nature of SR.



(a) Analysis of vanilla DMD in DBSR-to-DFSR distillation.

(b) Effect of the fidelity anchor on DBSR-to-DFSR distillation.

Figure 3: **Effect of fidelity anchoring on distribution matching.** (a) Vanilla DMD causes unstable optimization and rapid PSNR degradation when directly applied to SR. (b) Sample-level fidelity supervision stabilizes distribution matching and preserves reconstruction quality. The optimization landscapes are conceptual illustrations.

To this end, we propose **PhoenixSR**, a generative heterogeneous distillation framework that transfers a pretrained diffusion prior to independently designed diffusion-free SR networks through fidelity-aware distribution matching. During training, DFSR outputs are mapped to the diffusion latent space, where real and fake score models provide distribution-level guidance, while paired SR supervision preserves LR–HR correspondence. We introduce **Heterogeneous Distribution Adaptation (HDA)** to adapt score estimation to the SR training distribution and stabilize the evolving student distribution. We further employ **Directional Reliability Weighting (DRW)**, a lightweight sample reweighting strategy based on denoising-residual direction consistency, to reduce unstable score guidance. All diffusion-related components are removed after distillation, leaving the original DFSR architecture and inference cost unchanged.

The contributions are summarized as follows:

- To the best of our knowledge, PhoenixSR is the first framework that applies distribution matching distillation to transfer a pretrained diffusion generative prior into independently designed diffusion-free SR networks, without modifying their inference architecture.
- We develop a fidelity-aware distribution distillation scheme for paired SR. HDA adapts real- and fake-score estimation to the SR training distribution and combines distribution-level diffusion guidance with paired reconstruction supervision, while DRW further improves training stability through lightweight sample reweighting.
- Extensive experiments on three SR benchmarks and six CNN, Transformer, and lightweight SR backbones show that PhoenixSR consistently improves perceptual quality while largely preserving reconstruction fidelity and the original inference cost.

2 Related Work

We summarize the most relevant work here and defer detailed discussion to Appendix A. Real-world SR has evolved from CNN/Transformer reconstruction models (Dong et al., 2014; Kim et al., 2016; Lim et al., 2017; Zhang et al., 2018b; Liang et al., 2021; Chen et al., 2023) to diffusion-based methods that exploit strong generative priors (Wang et al., 2024a; Yue et al., 2023; Lin et al., 2024; Wu et al., 2024b; Yu et al., 2024), while efficient SR focuses on lightweight feed-forward inference (Kong et al., 2022; Sun et al., 2023; Zamfir et al., 2024). Conventional knowledge distillation transfers outputs or intermediate representations (Hinton et al., 2015; Romero et al., 2015; Heo et al., 2019; Gou et al., 2021), whereas most diffusion distillation methods accelerate sampling while retaining diffusion-derived generators (Salimans & Ho, 2022; Song et al., 2023; Meng et al., 2023). DMD/DMD2 instead perform distribution-level distillation through diffusion-score differences (Yin et al., 2024b;a), and related SDS/VSD methods also use pretrained diffusion scores as supervision (Poole et al., 2023; Wang et al., 2023b). Recent work has explored heterogeneous diffusion transfer in restoration: DTKD distills a diffusion SR teacher into a Transformer student through frequency-aware output supervision (Park & Song, 2026), DM-SR applies distribution matching to SR while retaining a diffusion denoiser at inference (Park & Lee, 2026), and DUO-VSR applies DMD to a diffusion-derived one-step VSR generator (Lv et al., 2026). Cross-space distillation further studies distribution transfer across different diffusion latent spaces (Nguyen et al., 2026). In contrast, PhoenixSR transfers a pretrained diffusion prior through distribution matching into independently designed diffusion-free SR networks, using paired supervision to preserve reconstruction fidelity and removing all diffusion components after training.

3 METHOD

Problem Formulation. Let $\mathbf{y}$ and $\mathbf{x}$ denote a degraded low-resolution (LR) image and its paired high-resolution (HR) target, respectively. A diffusion-free SR (DFSR) student G_θ predicts $\hat{\mathbf{x}}_0 = G_\theta(\mathbf{y})$. Since the diffusion teacher operates in Stable Diffusion latent space, its frozen VAE encoder E maps the student output to

$$\hat{\mathbf{x}}_0 = G_\theta(\mathbf{y}), \qquad \hat{\mathbf{z}}_0 = E(\hat{\mathbf{x}}_0), \qquad \hat{\mathbf{z}}_t = \alpha_t \hat{\mathbf{z}}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $\hat{\mathbf{z}}_t$ is the noisy student latent at timestep t. Our goal is to transfer the generative prior of a pretrained diffusion teacher to an independently designed DFSR student while preserving its original feed-forward inference architecture.

Following Distribution Matching Distillation (DMD) (Yin et al., 2024b;a), a real-score model characterizes the target distribution, while an online fake-score model tracks the evolving student distribution. In score notation, the distribution-matching gradient is

$$\nabla_\theta \mathcal{L}_{\mathrm{DMD}} = \mathbb{E}\left[w(t) \big(\mathbf{s}_{\mathrm{fake}}(\hat{\mathbf{z}}_t, t, \mathbf{c}) - \mathbf{s}_{\mathrm{real}}(\hat{\mathbf{z}}_t, t, \mathbf{c}) \big) \frac{\partial \hat{\mathbf{z}}_0}{\partial \theta} \right], \tag{2}$$

where $\mathbf{c}$ denotes the diffusion text condition and $w(t)$ absorbs timestep-dependent factors. The reported setting uses a fixed quality prompt, and the implementation realizes the score difference through the equivalent denoised-latent parameterization with per-sample normalization for stability. Importantly, DMD provides distribution-level supervision without requiring architectural correspondence between the diffusion model and the student, making it suitable for transferring diffusion priors to independently designed DFSR networks. However, distribution matching alone does not enforce consistency with the paired HR target. Preliminaries are provided in the supplementary material.

3.1 MOTIVATION

The architecture-independent nature of DMD provides a natural way to transfer diffusion generative priors to native DFSR networks. However, applying it to paired SR introduces two practical challenges. First, the real-score model should provide suitable guidance for the SR training domain, while the online fake-score model must continuously track the evolving student distribution. Second, unlike unconditional generation, SR requires each output to remain faithful to its corresponding LR–HR pair. As shown in Fig. 3, vanilla DMD causes unstable optimization and substantial PSNR degradation, whereas paired reconstruction supervision stabilizes the transfer. We therefore introduce **Heterogeneous Distribution Adaptation (HDA)** for score adaptation and fidelity anchoring, together with **Directional Reliability Weighting (DRW)** as an auxiliary sample reweighting strategy.

3.2 HETEROGENEOUS DISTRIBUTION ADAPTATION

Standard DMD uses a fixed real-score model for the target distribution and an online fake-score model for the generator distribution. Under diffusion-to-DFSR transfer, the real-score model needs to better match the SR training domain, while the fake-score model must track a continuously changing student distribution. Moreover, distribution matching alone does not guarantee paired reconstruction fidelity. HDA addresses these issues with an adapted real-score model, a representation-regularized fake-score model, and a sample-level fidelity anchor.

LoRA-adapted real-score model. The pretrained diffusion teacher learns a general natural-image prior, while the HR images used for real-world SR exhibit task-specific image statistics. Directly using the frozen score model may therefore provide suboptimal guidance for SR distillation.

To improve target-score compatibility while preserving the pretrained prior, we insert lightweight LoRA adapters (Hu et al., 2022) into the real-score model and freeze its original backbone. Given an HR target $\mathbf{x}$, let $\mathbf{z}_0^{HR} = E(\mathbf{x})$ and

$$\mathbf{z}_t^{HR} = \alpha_t \mathbf{z}_0^{HR} + \sigma_t \boldsymbol{\epsilon}, \qquad \mathcal{L}_{\mathrm{real}} = \mathbb{E}\left[\left\| \boldsymbol{\epsilon}_{\mathrm{real}}(\mathbf{z}_t^{HR}, t, \mathbf{c}) - \boldsymbol{\epsilon} \right\|_2^2 \right]. \tag{3}$$

Only the low-rank adapters are optimized, limiting adaptation while retaining the pretrained backbone. Detailed LoRA formulation and analysis are provided in the supplementary material.

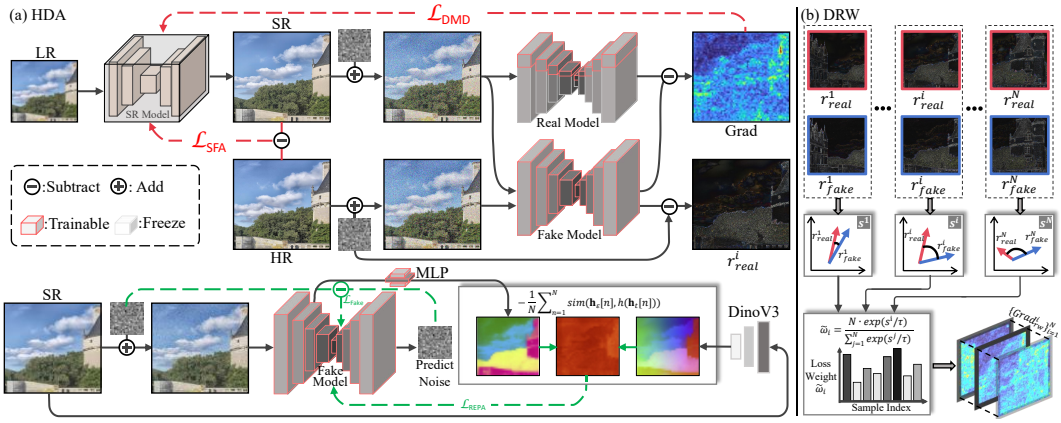


Figure 4: **Overview of PhoenixSR.** HDA adapts real-score estimation, tracks the evolving student distribution, and anchors paired fidelity. DRW reweights sample-wise guidance according to residual-direction consistency.

Representation-regularized fake-score model. The fake-score model estimates the evolving student distribution and is fully optimized online using noisy student outputs. Since the student output distribution changes continuously during distillation, directly fitting the fake-score model may make it sensitive to transient artifacts. We therefore regularize its intermediate representations following REPA (Yu et al., 2025), aligning fake-UNet features with a frozen DINOv3 encoder (Siméoni et al., 2026):

$$\mathcal{L}_{\text{fake}} = \mathbb{E}\left[\|\boldsymbol{\epsilon}_{\text{fake}}(\hat{\mathbf{z}}_t, t, \mathbf{c}) - \boldsymbol{\epsilon}\|_2^2\right], \qquad \mathcal{L}_{\text{fake}}^{\text{total}} = \mathcal{L}_{\text{fake}} + \lambda_{\text{repa}}\mathcal{L}_{\text{REPA}}. \tag{4}$$

This regularizes the fake-score representation while tracking the evolving student distribution. The full REPA objective is deferred to the supplementary material.

Sample-level fidelity anchor. Reliable score estimation alone is insufficient because DMD remains a distribution-level objective. For paired SR, the student should approach a high-quality image distribution while remaining faithful to the specific LR–HR pair. We therefore introduce a sample-level fidelity anchor (SFA):

$$\mathcal{L}_{\text{SFA}} = \lambda_{\text{pix}}\mathcal{L}_{\text{pix}} + \lambda_{\text{per}}\mathcal{L}_{\text{per}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}, \qquad \nabla_\theta \mathcal{L}_G = \lambda_{\text{dmd}}\nabla_\theta \mathcal{L}_{\text{DMD}} + \nabla_\theta \mathcal{L}_{\text{SFA}}. \tag{5}$$

The pixel and perceptual terms preserve reconstruction fidelity, while the adversarial term promotes realistic textures. As illustrated in Fig. 3(b), SFA constrains distribution matching with paired reconstruction supervision and substantially stabilizes training.

Together, LoRA adaptation improves target-score compatibility, representation-regularized online optimization improves student-distribution tracking, and SFA preserves input-specific reconstruction. These components enable distribution-level diffusion distillation to operate effectively on independently designed DFSR students.

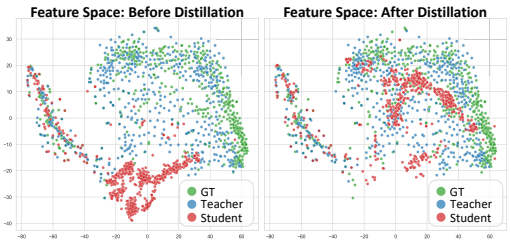


Figure 5: **t-SNE projection of feature representations.** After distillation, DFSR representations move toward the region between real HR samples and DBSR outputs in t-SNE projection.

As a qualitative representation-space diagnostic, Fig. 5 visualizes DFSR outputs before and after distillation together with real HR samples and DBSR outputs using t-SNE (van der Maaten & Hinton, 2008). After distillation, the projected DFSR representations move toward the intermediate region between real HR and DBSR outputs rather than directly overlapping the DBSR cluster, consistent with transferring diffusion priors while retaining HR-oriented reconstruction characteristics.

3.3 DIRECTIONAL RELIABILITY WEIGHTING

Although HDA improves real- and fake-score estimation, the online fake-score model may behave differently across samples as the student distribution continuously evolves. Samples that are not yet well represented by the fake-score model can therefore produce less stable distributional guidance. We use **Directional Reliability Weighting (DRW)** to adjust their relative contributions during training.

For each pair, the generated output and HR target are encoded as $\mathbf{z}_0^{G,i}$ and $\mathbf{z}_0^{HR,i}$ and perturbed using the same timestep t and Gaussian noise $\boldsymbol{\epsilon}^i$:

$$\mathbf{z}_t^{G,i} = \alpha_t \mathbf{z}_0^{G,i} + \sigma_t \boldsymbol{\epsilon}^i, \qquad \mathbf{z}_t^{HR,i} = \alpha_t \mathbf{z}_0^{HR,i} + \sigma_t \boldsymbol{\epsilon}^i. \tag{6}$$

Using the same perturbation removes variation caused purely by timestep or sampled noise, allowing the fake-score responses around the generated and HR latents to be compared under the same condition.

We evaluate both noisy states using the fake-score model and compute their denoising residuals and directional consistency:

$$\mathbf{r}_G^i = \boldsymbol{\epsilon}_{\text{fake}}(\mathbf{z}_t^{G,i}, t, \mathbf{c}^i) - \boldsymbol{\epsilon}^i, \quad \mathbf{r}_{HR}^i = \boldsymbol{\epsilon}_{\text{fake}}(\mathbf{z}_t^{HR,i}, t, \mathbf{c}^i) - \boldsymbol{\epsilon}^i, \quad s_i = \cos(\mathbf{r}_G^i, \mathbf{r}_{HR}^i). \tag{7}$$

Here $\mathbf{c}^i$ denotes the diffusion condition. We use the residual-direction consistency as a sample-wise proxy for the local behavior of the fake-score model rather than as an image-quality measure. Higher s_i indicates more similar denoising behavior around the generated and HR regions.

We convert s_i into a normalized reliability weight using a temperature-controlled softmax and clip extreme values for stability:

$$\omega_i = \text{clip}\left(N \frac{\exp(s_i/\tau)}{\sum_{j=1}^{N} \exp(s_j/\tau)}, \omega_{\min}, \omega_{\max} \right), \qquad \mathcal{L}_{\text{DMD}}^{\text{DRW}} = \frac{1}{N} \sum_{i=1}^{N} \text{sg}(\omega_i) \mathcal{L}_{\text{DMD}}^i. \tag{8}$$

Here N is the number of samples in the reliability-normalization pool; in distributed training, similarities are gathered across workers before the softmax. The factor N places the softmax weights on a unit-mean scale before clipping, while τ controls the contrast among samples. Here $\text{sg}(\cdot)$ denotes stop-gradient; no gradient is propagated through s_i or ω_i during the generator update. The resulting weights are used to adjust the relative contribution of different samples to the DMD objective.

The timestep schedule, teacher warm-up, asynchronous optimization, and complete PhoenixSR training procedure are provided in the supplementary material.

4 EXPERIMENTS

Implementation. Following the real-world SR setting of SwinIR (Liang et al., 2021), we construct the HR training set from DIV2K (Agustsson & Timofte, 2017), Flickr2K (Lim et al., 2017), OST (Wang et al., 2018), WED (Ma et al., 2017), FFHQ (Karras et al., 2019), Manga109 (Matsui et al., 2017), and SCUT-CTW1500 (Liu et al., 2019), with LR inputs synthesized using the Real-ESRGAN degradation pipeline (Wang et al., 2021). We evaluate $4\times$ SR on RealSR (Cai et al., 2019), DRealSR (Wei et al., 2020), and DIV2K-Val (Agustsson & Timofte, 2017), using PSNR (Hore & Ziou, 2010), LPIPS (Zhang et al., 2018a), MANIQA (Yang et al., 2022), NIQE (Mittal et al., 2013), MUSIQ (Ke et al., 2021), CLIP-IQA (Wang et al., 2023a), LIQE-Mix (Zhang et al., 2023), ARNIQA (Agnolucci et al., 2024), and TOPIQ (Chen et al., 2024). PhoenixSR is applied to SwinIR-L/M/S (Liang et al., 2021), HAT (Chen et al., 2023), Real-ESRGAN (Wang et al., 2021), and SeeMoRe (Zamfir et al., 2024), using pretrained Stable Diffusion 2.1 (Rombach et al., 2022) as the training-only diffusion teacher with a fixed quality prompt; the real-score model is LoRA-adapted (Hu et al., 2022), while the fake-score model is optimized online with REPA regularization (Yu et al., 2025). Quantitative comparisons include StableSR (Wang et al., 2024a), SeeSR (Wu et al., 2024b), SinSR (Wang et al., 2024b), OSEDiff (Wu et al., 2024a), and AdcSR (Chen et al., 2025), with additional diffusion-based methods shown qualitatively. All auxiliary diffusion components are discarded after distillation, leaving the student inference architecture unchanged; detailed optimization and hyperparameter settings are provided in the supplementary material.

Table 1: Quantitative comparison on RealSR. PhoenixSR is applied to DFSR backbones, while DBSR methods are included as reference baselines. Gray cells denote results with PhoenixSR; red and blue denote performance improvement and degradation, respectively.

Model	PSNR↑	LPIPS↓	MANIQA↑	NIQE↓	MUSIQ↑	CLIP-IQA↑	LIQE-Mix↑	ARNIQA↑	TOPIQ↑	Params (M)	FLOPs (G)	Runtime (ms)
DBSR Models												
StableSR	24.4272	0.2636	0.5910	6.5779	61.8041	0.5644	3.5794	0.6772	0.5121	1410	162731	17870
SeeSR	25.1475	0.3007	0.6451	5.3958	69.8181	0.6703	3.6576	0.7152	0.6888	2511	131980	5451
SinSR	26.1309	0.3068	0.5432	6.1038	61.2905	0.6236	3.5595	0.6601	0.5309	119	5392	201
OSEDiff	25.1554	0.2928	0.6343	5.6243	69.1793	0.6669	4.1382	0.6935	0.6292	1761	4611	169
AdcSR	24.4016	0.2939	0.6372	5.2979	70.0261	0.6939	4.2198	0.7227	0.6863	456	990	101
DFSR Models												
SwinIR-L	26.3071	0.2531	0.5228	5.7025	58.6388	0.4502	3.0465	0.6593	0.4774			
(w/PhoenixSR)	26.7437	0.2500	0.5508	5.4432	61.3691	0.4824	3.1735	0.6638	0.5063	28.01	1006.56	202
Gain	+0.4366	-0.0031	+0.0280	-0.2593	+2.7303	+0.0322	+0.1270	+0.0045	+0.0289			
HAT	25.9792	0.2421	0.5517	6.1520	57.9440	0.4068	3.1663	0.6552	0.4660			
(w/PhoenixSR)	26.1969	0.2399	0.5708	5.9102	61.1100	0.4599	3.4304	0.6732	0.5073	20.77	833.45	201
Gain	+0.2177	-0.0022	+0.0191	-0.2418	+3.1660	+0.0531	+0.2641	+0.0180	+0.0413			
Real-ESRGAN	25.6900	0.2708	0.5505	5.7937	60.3978	0.4761	3.3976	0.6754	0.5154			
(w/PhoenixSR)	25.9452	0.2626	0.5590	5.6776	61.6747	0.4740	3.6010	0.6827	0.5302	16.70	587.62	50
Gain	+0.2552	-0.0082	+0.0085	-0.1161	+1.2769	-0.0021	+0.2034	+0.0073	+0.0148			
SwinIR-M	25.8912	0.2608	0.5226	5.7635	59.1610	0.4435	3.1377	0.6634	0.4860			
(w/PhoenixSR)	26.3384	0.2593	0.5572	5.4158	62.1048	0.5112	3.2828	0.6730	0.5182	11.72	448.57	111
Gain	+0.4472	-0.0015	+0.0346	-0.3477	+2.9438	+0.0677	+0.1451	+0.0096	+0.0322			
SwinIR-S	25.3030	0.2965	0.4413	7.4007	54.2403	0.4463	2.7434	0.6502	0.4458			
(w/PhoenixSR)	25.2573	0.3035	0.5529	5.5930	65.3453	0.5790	3.9146	0.6957	0.5851	1.05	79.95	61
Gain	-0.0457	+0.0070	+0.1116	-1.8077	+11.1050	+0.1327	+1.1712	+0.0455	+0.1393			
SeeMoRe	26.0739	0.2953	0.3476	8.6255	48.7427	0.3813	1.9931	0.4839	0.3151			
(w/PhoenixSR)	25.9329	0.2663	0.5454	5.1620	62.4174	0.5283	3.6825	0.6757	0.4976	0.97	28.55	40
Gain	-0.1410	-0.0290	+0.1978	-3.4635	+13.6747	+0.1470	+1.6894	+0.1918	+0.1825			

Table 2: Quantitative comparison on DRealSR.

Model	PSNR↑	LPIPS↓	MANIQA↑	NIQE↓	MUSIQ↑	CLIP-IQA↑	LIQE-Mix↑	ARNIQA↑	TOPIQ↑
DBSR Models									
StableSR	28.4070	0.2726	0.4914	7.5141	51.4058	0.5046	2.6971	0.5762	0.4321
SeeSR	28.0712	0.3174	0.6052	6.4061	65.0873	0.6909	3.6846	0.6880	0.6572
SinSR	28.2012	0.3511	0.5034	6.7220	57.5658	0.6574	3.5766	0.6224	0.5288
OSEDiff	27.9321	0.2971	0.5874	6.5227	64.8184	0.6909	3.9768	0.6702	0.6005
AdcSR	26.2041	0.3156	0.5937	6.1738	66.8444	0.7206	4.2538	0.6985	0.6602
DFSR Models									
SwinIR-L	28.4999	0.2739	0.4751	6.6093	52.8383	0.4680	2.9287	0.5935	0.4425
(w/PhoenixSR)	29.1477	0.2721	0.5095	6.1132	54.6482	0.5006	3.1093	0.6045	0.4663
Gain	+0.6478	-0.0018	+0.0344	-0.4961	+1.8099	+0.0326	+0.1806	+0.0110	+0.0238
HAT	28.9684	0.2594	0.4887	7.1056	50.8062	0.4196	2.8382	0.5710	0.4183
(w/PhoenixSR)	29.0597	0.2546	0.5134	6.4380	54.1864	0.4735	3.1686	0.6105	0.4510
Gain	+0.0913	-0.0048	+0.0247	-0.6676	+3.3802	+0.0539	+0.3304	+0.0395	+0.0327
Real-ESRGAN	28.6276	0.2819	0.4903	6.7022	54.2878	0.4812	3.1287	0.6103	0.4627
(w/PhoenixSR)	28.7265	0.2790	0.5041	6.3026	55.8058	0.5072	3.3049	0.6282	0.4771
Gain	+0.0989	-0.0029	+0.0138	-0.3996	+1.5180	+0.0260	+0.1762	+0.0179	+0.0144
SwinIR-M	28.1926	0.2827	0.4672	6.5990	53.1293	0.4517	2.9558	0.6054	0.4518
(w/PhoenixSR)	28.7346	0.2811	0.5076	5.9858	55.3788	0.5172	3.1711	0.6248	0.4793
Gain	+0.5420	-0.0016	+0.0404	-0.6132	+2.2495	+0.0655	+0.2153	+0.0194	+0.0275
SwinIR-S	28.5549	0.3044	0.4065	8.5537	49.5158	0.4196	2.5339	0.5988	0.4223
(w/PhoenixSR)	28.3787	0.3125	0.4985	6.3289	59.6194	0.5719	3.6983	0.6641	0.5341
Gain	-0.1762	+0.0081	+0.0920	-2.2248	+10.1036	+0.1523	+1.1644	+0.0653	+0.1118
SeeMoRe	28.9936	0.3264	0.2894	10.3603	39.9581	0.4225	1.6149	0.3754	0.2818
(w/PhoenixSR)	28.8592	0.2818	0.5060	6.1061	56.2313	0.5038	3.3956	0.6304	0.4696
Gain	-0.1344	-0.0446	+0.2166	-4.2542	+16.2732	+0.0813	+1.7807	+0.2550	+0.1878

4.1 Quantitative Comparison and Qualitative Results

Tables 1, 2, and 3 report quantitative results on RealSR, DRealSR, and DIV2K-Val, respectively. Across six CNN, Transformer, and lightweight DFSR backbones, PhoenixSR consistently improves perceptual quality while largely preserving reconstruction fidelity. For medium and large models, the improvement is generally well balanced. On RealSR, SwinIR-L, HAT, Real-ESRGAN, and SwinIR-M improve PSNR by 0.44, 0.22, 0.26, and 0.45 dB, respectively, together with improvements on most perceptual metrics. Similar trends are observed on DRealSR, where SwinIR-L and SwinIR-M gain 0.65 and 0.54 dB in PSNR while improving all reported no-reference metrics.

The gains are particularly noticeable for lightweight students. SwinIR-S (1.05M) and SeeMoRe (0.97M) remain within approximately 0.2 dB of their original PSNR across the three datasets, while obtaining substantially better perceptual scores. For example, PhoenixSR improves MUSIQ by 11.11/10.10/17.73 for SwinIR-S and 13.67/16.27/23.33 for SeeMoRe on RealSR/DRealSR/DIV2K-Val, respectively. SeeMoRe further improves LPIPS on all three benchmarks. These results suggest that distribution-level diffusion guidance can provide useful generative priors even for highly compact feed-forward SR networks.

Compared with DBSR baselines, PhoenixSR targets a different inference regime. StableSR, SeeSR, OSEDiff, and AdcSR retain diffusion-based or diffusion-derived generation pipelines at inference,

Table 3: Quantitative comparison on DIV2K-Val.

Model	PSNR↑	LPIPS↓	MANIQA↑	NIQE↓	MUSIQ↑	CLIP-IQA↑	LIQE-Mix↑	ARNIQA↑	TOPIQ↑
DBSR Models									
StableSR	24.2395	0.3129	0.5506	5.6054	58.4136	0.5627	3.4477	0.6580	0.4851
SeeSR	23.6780	0.3194	0.6222	4.8095	68.6710	0.6935	4.0053	0.7301	0.6853
SinSR	24.2883	0.3222	0.5414	5.8005	63.2651	0.6530	3.7998	0.6745	0.5796
OSEDiff	23.7214	0.2942	0.6130	4.7203	67.9926	0.6611	4.2976	0.7030	0.6194
AdcSR	23.3567	0.2845	0.6071	4.3420	68.0922	0.6846	4.3009	0.7215	0.6574
DFSR Models									
SwinIR-L	23.9606	0.3152	0.5402	4.7492	60.1092	0.5246	3.5484	0.6607	0.5071
(w/PhoenixSR)	24.1746	0.2996	0.5675	4.3867	62.4468	0.5793	3.8263	0.6760	0.5372
Gain	+0.2140	-0.0156	+0.0273	-0.3625	+2.3376	+0.0547	+0.2779	+0.0153	+0.0301
HAT	24.7455	0.2931	0.5547	5.1686	58.2528	0.4622	3.4201	0.6467	0.4742
(w/PhoenixSR)	24.3991	0.2831	0.5853	4.6309	62.6893	0.5514	3.9650	0.6824	0.5446
Gain	-0.3464	-0.0100	+0.0306	-0.5377	+4.4365	+0.0892	+0.5449	+0.0357	+0.0704
Real-ESRGAN	24.2963	0.3116	0.5484	4.6792	61.0717	0.5274	3.6659	0.6738	0.5300
(w/PhoenixSR)	24.2379	0.3112	0.5437	4.3980	62.1329	0.5569	3.7818	0.6841	0.5378
Gain	-0.0584	-0.0004	-0.0047	-0.2812	+1.0612	+0.0295	+0.1159	+0.0103	+0.0078
SwinIR-M	23.9014	0.3235	0.5164	4.7827	58.1455	0.5170	3.4387	0.6513	0.4838
(w/PhoenixSR)	24.1663	0.3089	0.5477	4.4839	61.5357	0.5842	3.7218	0.6774	0.5226
Gain	+0.2649	-0.0146	+0.0313	-0.2988	+3.3902	+0.0672	+0.2831	+0.0261	+0.0388
SwinIR-S	23.9349	0.3764	0.3965	7.0025	46.0279	0.3890	2.0897	0.5831	0.3759
(w/PhoenixSR)	23.7942	0.3390	0.5176	4.2790	63.7584	0.6107	3.8624	0.6948	0.5674
Gain	-0.1407	-0.0374	+0.1211	-2.7235	+17.7305	+0.2217	+1.7727	+0.1117	+0.1915
SeeMoRe	24.8031	0.4907	0.2661	8.9587	37.7098	0.3919	1.5603	0.4023	0.2374
(w/PhoenixSR)	24.6089	0.3116	0.5164	4.2905	61.0420	0.5363	3.7589	0.6702	0.4916
Gain	-0.1942	-0.1791	+0.2503	-4.6682	+23.3322	+0.1444	+2.1986	+0.2679	+0.2542

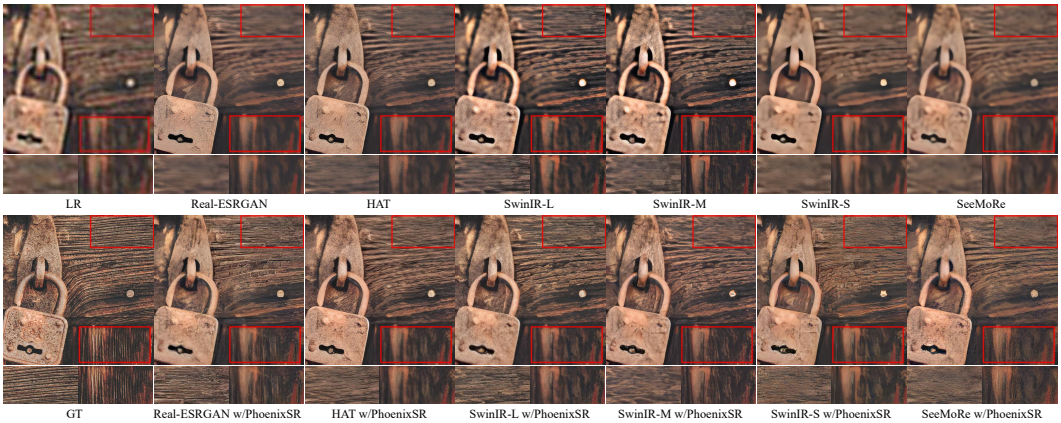


Figure 6: **Qualitative comparison on DIV2K-Val.** PhoenixSR enhances high-frequency details of diffusion-free SR models, producing clearer and more natural wood textures while preserving the underlying image structure.

whereas all diffusion-related components in PhoenixSR are removed after distillation. As shown in Table 1, the distilled students preserve the parameter count, FLOPs, and runtime of their original backbones. Figures 6 and 7 further show that PhoenixSR improves local textures while largely preserving the underlying image structure. The consistent behavior across different student architectures indicates that the proposed diffusion-prior transfer is not restricted to a particular DFSR backbone.

4.2 Ablation Studies

We evaluate diffusion-transfer strategies and the contributions of PhoenixSR components. The tables report results from the full training schedule, whereas controlled comparisons and ablations use a fixed 80K-iteration budget with the same number of student updates. Their absolute values are not directly comparable, as different training stages may correspond to different perception–distortion operating points. All ablation conclusions are drawn only within the same controlled setting.

Comparison with alternative transfer strategies. We compare PhoenixSR with strong task-only fine-tuning, output-level heterogeneous distillation, SDS, and DMD2-based distribution matching. All methods start from the same student checkpoint and use the same training data, number of student updates, paired SR supervision when applicable, and validation-based model selection protocol.

Table 4 shows that PhoenixSR achieves a better fidelity–perception balance than task-only fine-tuning, output KD, SDS, and DMD2-based transfer across the three representative students. PhoenixSR



Figure 7: **Qualitative comparison with diffusion-based SR methods.** Diffusion-based baselines may over-sharpen the statue surface or introduce hallucinated textures that deviate from the underlying geometry. PhoenixSR improves details while retaining more faithful structural reconstruction.

Table 4: Comparison of alternative diffusion-to-DFSR transfer strategies on RealSR.

Method	SwinIR-L			SwinIR-S			SeeMoRe		
	PSNR↑	LPIPS↓	MUSIQ↑	PSNR↑	LPIPS↓	MUSIQ↑	PSNR↑	LPIPS↓	MUSIQ↑
Task-only strong FT	26.41	0.257	59.7	25.13	0.301	56.0	26.00	0.289	53.7
Output KD + GT	25.97	0.274	56.7	24.77	0.327	55.8	25.98	0.301	52.1
SDS + SFA	25.77	0.273	56.5	24.89	0.305	50.4	25.10	0.317	51.8
DMD2 + SFA	25.55	0.256	57.8	24.78	0.311	54.7	24.87	0.304	53.6
PhoenixSR w/o DRW	26.15	0.247	63.5	24.79	0.296	61.5	25.35	0.290	60.8
PhoenixSR	26.44	0.244	64.3	25.21	0.287	63.4	25.93	0.279	61.5

without DRW already provides strong perceptual gains, while DRW mainly improves reconstruction fidelity and the overall balance. These results indicate that the main benefit comes from the proposed fidelity-aware distribution-level transfer rather than the auxiliary reweighting strategy.

Component ablation. We further analyze the main components on SwinIR-L. "Score Adapt." denotes LoRA real-score adaptation together with REPA-regularized online fake-score optimization, while SFA denotes the sample-level fidelity anchor. Table 5 shows that distribution matching without paired fidelity supervision causes severe degradation, confirming the importance of SFA for fidelity-sensitive SR. Combining SFA with score adaptation substantially improves perceptual quality, while DRW further improves PSNR and CLIP-IQA, yielding a better fidelity–perception balance. Additional ablations of score-model design, DRW, and the training schedule, together with robustness analyses, qualitative results, and the user study, are provided in the supplementary material.

Table 5: Ablation of the main PhoenixSR components.

Variant	SFA	Score Adapt.	DRW	PSNR↑	MANIQA↑	MUSIQ↑	CLIP-IQA↑
Base SwinIR-L	–	–	–	26.3071	0.5228	58.6388	0.4502
SFA only	✓	–	–	26.3124	0.5067	56.8245	0.4538
w/o SFA	–	✓	✓	20.7107	0.3512	40.7941	0.3071
SFA + Score Adapt.	✓	✓	–	26.1587	0.5678	63.5729	0.4963
Full PhoenixSR	✓	✓	✓	26.4409	0.5745	64.3102	0.5026

5 Conclusion

We propose **PhoenixSR**, a generative heterogeneous distillation framework that transfers pretrained diffusion priors to independently designed diffusion-free SR networks through score-based distribution matching. Paired SR supervision preserves reconstruction fidelity, while HDA adapts score estimation to the SR setting and DRW further improves the fidelity–perception balance. Experiments across six feed-forward SR backbones and multiple real-world benchmarks demonstrate consistent perceptual improvements without changing the student inference architecture. Controlled comparisons with alternative transfer strategies further support the effectiveness of distribution-level diffusion-prior transfer for efficient SR.

REPRODUCIBILITY STATEMENT

The main paper specifies the model components, objectives, datasets, evaluation metrics, and deployment costs. Detailed optimization settings, training procedure, and additional ablations are provided in the supplementary material to support reproduction of the reported results. Anonymous code can be provided to reviewers upon request during the discussion phase.

ETHICS STATEMENT

The user study involved 20 adult participants and was conducted solely for subjective image-quality evaluation. All participants were informed of the purpose and procedure of the study and participated voluntarily. No personally identifiable or sensitive information was collected, and participants received no monetary or other compensation. Under our institution's policy, this anonymous, non-medical, low-risk image-rating study does not require formal ethics-board approval.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript language editing, organization, literature/reference checking, and consistency checking between the manuscript and implementation. No measured experimental data or quantitative results were generated by these tools. All AI-assisted content was reviewed and verified by the authors, who take responsibility for the final manuscript.

REFERENCES

Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. ARNIQA: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 189–198, 2024.

Eirikur Agustsson and Radu Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 126–135, 2017.

Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision*, pp. 252–268, 2018.

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

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22367–22377, 2023.

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

Bin Chen, Gehui Li, Rongyuan Wu, Xindong Zhang, Jie Chen, Jian Zhang, and Lei Zhang. Adversarial diffusion compression for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28208–28220, 2025.

Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11065–11074, 2019.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pp. 184–199, 2014.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2): 295–307, 2016a.

Chao Dong, Chen Change Loy, and Xiaoou Tang. Accelerating the super-resolution convolutional neural network. In *Proceedings of the European Conference on Computer Vision*, pp. 391–407, 2016b.

Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. TSD-SR: One-Step Diffusion with Target Score Distillation for Real-World Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23174–23184, 2025.

Jianping Gou, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

Jinjin Gu, Hannan Lu, Wangmeng Zuo, and Chao Dong. Blind super-resolution with iterative kernel correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1604–1613, 2019.

Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1921–1930, 2019.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Alain Hore and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *Proceedings of the International Conference on Pattern Recognition*, pp. 2366–2369, 2010.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, 2022.

Zheng Hui, Xiumei Wang, and Xinbo Gao. Fast and accurate single image super-resolution via information distillation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 723–731, 2018.

Zheng Hui, Xinbo Gao, Yunchu Yang, and Xiumei Wang. Lightweight image super-resolution with information multi-distillation network. In *Proceedings of the ACM International Conference on Multimedia*, pp. 2024–2032, 2019.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, pp. 694–711, 2016.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

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

Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1646–1654, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

Fangyuan Kong, Mingxi Li, Songwei Liu, Ding Liu, Jingwen He, Yang Bai, Fangmin Chen, and Lean Fu. Residual local feature network for efficient super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 766–776, 2022.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the International Conference on Machine Learning*, volume 97, pp. 3519–3529, 2019.

Wenbo Li, Kun Zhou, Lu Qi, Nianjuan Jiang, Jiangbo Lu, and Jiaya Jia. LAPAR: Linearly-assembled pixel-adaptive regression network for single image super-resolution and beyond. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20343–20355, 2020.

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 Workshops*, pp. 1833–1844, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136–144, 2017.

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 *Proceedings of the European Conference on Computer Vision*, pp. 430–448, 2024.

Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition*, 90:337–345, 2019. doi: 10.1016/j.patcog.2019.02.002

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*, 2019.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5775–5787, 2022.

Zhengyao Lv, Menghan Xia, Xintao Wang, and Kwan-Yee K. Wong. DUO-VSR: Dual-Stream Distillation for One-Step Video Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16333–16344, 2026.

Kede Ma, Zhengfang Duanmu, Qingbo Wu, Zhou Wang, Hongwei Yong, Hongliang Li, and Lei Zhang. Waterloo Exploration Database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, 26(2):1004–1016, 2017.

Yusuke Matsui, Kota Ito, Yuji Aramaki, Azuma Fujimoto, Toru Ogawa, Toshihiko Yamasaki, and Kiyoharu Aizawa. Sketch-based Manga retrieval using Manga109 dataset. *Multimedia Tools and Applications*, 76(20):21811–21838, 2017.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14297–14306, 2023.

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

Anh Nguyen, Ngan Nguyen, Duc Vu, Trung Dao, Viet Nguyen, Quan Dao, Kien Nguyen, Chi Tran, Phong Nguyen, Khoi Nguyen, Cuong Pham, Dimitris Metaxas, Vishal M. Patel, and Anh Tran. Cross-Space Distillation: Teaching One-Step Students with Modern Diffusion Teachers. *arXiv preprint arXiv:2606.32020*, 2026.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171, 2021.

Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *Proceedings of the European Conference on Computer Vision*, pp. 191–207, 2020.

Joonkyu Park and Kyoung Mu Lee. Bridging the Distribution Gap to Harness Pretrained Diffusion Priors for Super-Resolution. In *International Conference on Learning Representations*, 2026.

Jeong Hyeok Park and Byung Cheol Song. DTKD: Diffusion-to-Transformer Heterogeneous Knowledge Distillation for Efficient and Perceptually Enhanced Super-Resolution. *Electronics*, 15(10): 1986, 2026. doi: 10.3390/electronics15101986

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *International Conference on Learning Representations*, 2023.

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

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

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

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

Daniil Selikhanovych, David Li, Aleksei Leonov, Nikita Gushchin, Sergei Kushneriuk, Alexander Filippov, Evgeny Burnaev, Iaroslav Koshelev, and Alexander Korotin. One-Step Residual Shifting Diffusion for Image Super-Resolution via Distillation. In *International Conference on Machine Learning*, 2026.

Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883, 2016.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pp. 32211–32252, 2023.

Long Sun, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Spatially-adaptive feature modulation for efficient image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13190–13199, 2023.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 606–615, 2018.

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

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023a. No. 2.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. ProlificDreamer: High-fidelity and diverse text-to-3D generation with variational score distillation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8406–8441, 2023b.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132:5929–5949, 2024a.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. SinSR: Diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25796–25805, 2024b.

Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. Component divide-and-conquer for real-world image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pp. 101–117, 2020.

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

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

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *Proceedings of the European Conference on Computer Vision*, pp. 74–91, 2024.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems*, volume 37, pp. 47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6613–6623, 2024b.

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

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *Proceedings of the International Conference on Learning Representations*, 2025.

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

Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yulun Zhang, and Radu Timofte. See more details: Efficient image super-resolution by experts mining. In *Proceedings of the International Conference on Machine Learning*, volume 235, pp. 58158–58173, 2024.

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

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision*, pp. 286–301, 2018b.

Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *International Conference on Learning Representations*, 2019.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4791–4800, 2021.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14071–14081, 2023.

Supplementary Material

A Extended Related Work

A.1 Real-World Image Super-Resolution

Single-image super-resolution has progressed from CNN-based reconstruction models, including SRCNN (Dong et al., 2014), VDSR (Kim et al., 2016), EDSR (Lim et al., 2017), and RCAN (Zhang et al., 2018b), to Transformer architectures such as SwinIR (Liang et al., 2021) and HAT (Chen et al., 2023). While these methods perform strongly under synthetic degradations, real-world SR additionally requires handling unknown blur, noise, compression, and camera-processing artifacts. Blind and real-world SR methods therefore estimate degradation kernels (Gu et al., 2019) or construct practical degradation pipelines for synthesizing realistic LR–HR pairs (Zhang et al., 2021; Wang et al., 2021). RealSR (Cai et al., 2019) and DRealSR (Wei et al., 2020) provide real-captured evaluation settings.

Diffusion models provide stronger generative priors for perceptual restoration. SR3 (Saharia et al., 2023) formulates SR as iterative denoising, while latent diffusion (Rombach et al., 2022) enables scalable pretrained generative priors. StableSR (Wang et al., 2024a), ResShift (Yue et al., 2023), DiffBIR (Lin et al., 2024), SeeSR (Wu et al., 2024b), PASD (Yang et al., 2024), and SUPIR (Yu et al., 2024) exploit such priors for real-world restoration. SinSR (Wang et al., 2024b), OSEDiff (Wu et al., 2024a), and AdcSR (Chen et al., 2025) reduce the number of diffusion steps, but still retain diffusion-derived generation pipelines at inference.

DM-SR (Park & Lee, 2026) further introduces distribution matching into SR by mapping degraded inputs toward distributions compatible with a pretrained diffusion model, while the diffusion denoiser remains part of inference. PhoenixSR instead uses the pretrained diffusion model only as training-time distributional supervision and removes all diffusion components after distillation.

A.2 Lightweight and Efficient Super-Resolution

Efficient SR is important for mobile, embedded, and real-time applications. Early methods reduce computation by operating mainly in LR feature space and using efficient upsampling (Shi et al., 2016; Dong et al., 2016b). CARN (Ahn et al., 2018) improves feature reuse through cascading residual connections, while IDN (Hui et al., 2018) and IMDN (Hui et al., 2019) introduce information-distillation mechanisms for compact feature representations. RLFN (Kong et al., 2022), LAPAR (Li et al., 2020), and SAFMN (Sun et al., 2023) further improve the efficiency–accuracy trade-off through lightweight feature aggregation, pixel-adaptive regression, or spatial feature modulation.

Transformer-based restoration models provide another family of feed-forward SR networks. SwinIR (Liang et al., 2021) introduces shifted-window attention for image restoration, while HAT (Chen et al., 2023) combines window and channel attention for stronger reconstruction capability. SeeMoRe (Zamfir et al., 2024) achieves efficient detail restoration with approximately one million parameters through expert-based feature modeling.

A closely related direction is heterogeneous knowledge transfer from diffusion models to efficient SR networks. DTKD (Park & Song, 2026) transfers perceptual knowledge from a diffusion SR teacher to a Transformer student through frequency-aware output distillation. PhoenixSR shares the goal of diffusion-to-feed-forward knowledge transfer, but transfers the pretrained diffusion prior through score-based distribution matching rather than directly matching sampled teacher outputs.

A.3 Knowledge Distillation and Diffusion Distillation

Knowledge distillation transfers knowledge from a high-capacity teacher to a smaller student (Hinton et al., 2015; Gou et al., 2021). Beyond response-level supervision, later methods align intermediate features (Romero et al., 2015; Heo et al., 2019), similarity structures (Tung & Mori, 2019), and relational information (Park et al., 2019). Such approaches are most natural when teacher and student possess compatible representations or deterministic output correspondence.

Table 6: **Positioning of PhoenixSR with closely related diffusion-transfer methods.** "Native DFSR" denotes a conventional feed-forward SR architecture designed independently from the diffusion teacher. "Diffusion-derived inference" indicates whether the deployed model still relies on a diffusion model or a diffusion-derived generator.

Method	Task	Transfer mechanism	Deployed model	Native DFSR	Diffusion-derived inference
DTKD (Park & Song, 2026)	SR	Frequency-aware output KD	Transformer SR student	✓	No
DM-SR (Park & Lee, 2026)	SR	Distribution alignment	Pretrained diffusion model with learned input mapping	–	Yes
Cross-Space Distillation (Nguyen et al., 2026)	Image generation	Distribution-level diffusion distillation	One-step diffusion student	–	Yes
PhoenixSR	SR	Score-based distribution matching	CNN / Transformer / lightweight SR students	✓	No

Diffusion distillation primarily accelerates multi-step diffusion generators. Progressive distillation (Salimans & Ho, 2022), consistency models (Song et al., 2023), and guided diffusion distillation (Meng et al., 2023) reduce the number of denoising steps while retaining diffusion-derived generators. Distribution Matching Distillation (DMD) (Yin et al., 2024b) combines distribution matching through real–fake score differences with a teacher-output regression loss for stable training. DMD2 (Yin et al., 2024a) removes the regression requirement using two-time-scale updates and additionally introduces real-data adversarial supervision and a training procedure for few-step generation. Related score-distillation approaches such as SDS (Poole et al., 2023) and VSD (Wang et al., 2023b) also demonstrate the effectiveness of pretrained diffusion scores as distribution-level supervision.

DMD has also been extended to video SR. DUO-VSR (Lv et al., 2026) combines distribution matching with additional supervision to obtain one-step video SR, but its generator remains diffusion-derived. In contrast, PhoenixSR transfers a pretrained diffusion prior to independently designed diffusion-free SR students and preserves their original feed-forward inference architectures.

A.4 Positioning with Closely Related Diffusion-Transfer Methods

Several recent works are closely related to PhoenixSR in their use of diffusion priors for efficient generation or super-resolution. We therefore further clarify the specific setting addressed by PhoenixSR and its differences from DTKD (Park & Song, 2026), DM-SR (Park & Lee, 2026), RSD (Selikhanovych et al., 2026), TSD-SR (Dong et al., 2025), and Cross-Space Distillation (Nguyen et al., 2026).

The novelty of PhoenixSR does not simply lie in transferring knowledge from a diffusion model to an efficient SR network. DTKD has already demonstrated diffusion-to-Transformer heterogeneous distillation while retaining feed-forward student inference. Instead, PhoenixSR focuses on transferring a pretrained diffusion prior through *score-based distribution matching* into pre-existing, independently designed diffusion-free SR architectures, without retaining a diffusion-derived generator or modifying the native student inference graph.

Difference from DTKD. DTKD (Park & Song, 2026) transfers perceptual knowledge from a diffusion SR teacher to a Transformer student through output-level supervision. It decomposes teacher–student output discrepancies into frequency sub-bands and emphasizes selected frequency components during distillation. PhoenixSR shares the broader goal of diffusion-to-feed-forward SR transfer, but adopts a different form of supervision. Rather than requiring the student to imitate a sampled diffusion output, PhoenixSR uses the pretrained diffusion model to provide distribution-level guidance through real–fake score differences. This avoids enforcing direct sample-wise correspondence with a particular diffusion reconstruction, which may contain perceptually plausible details that do not exactly agree with the paired HR target.

PhoenixSR is also evaluated across CNN, Transformer, and lightweight SR architectures. These students are independently designed feed-forward SR networks rather than architectures derived from the diffusion teacher, and their original inference graphs remain unchanged after distillation.

Difference from DM-SR. DM-SR (Park & Lee, 2026) also introduces distribution matching into super-resolution, but addresses a different problem. It learns an encoder that maps degraded LR observations toward a distribution compatible with a pretrained diffusion model, after which the pretrained diffusion model remains responsible for restoration at inference time.

PhoenixSR instead performs distribution matching directly on an independently designed diffusion-free SR student. The diffusion model is used only to provide training-time supervision. After distillation, the diffusion score models, VAE, and other auxiliary components are removed, and the

deployed model is exactly the original feed-forward SR student. Therefore, PhoenixSR transfers the diffusion prior into the student itself rather than preserving the diffusion model as part of the restoration pipeline.

Difference from RSD and TSD-SR. RSD (Selikhanovych et al., 2026) combines distribution-based ResShift distillation with ground-truth perceptual supervision and an adversarial loss. TSD-SR (Dong et al., 2025) incorporates real-image reference scores and reconstruction supervision into one-step diffusion SR. Both retain diffusion-derived generators at inference and already combine diffusion guidance with restoration supervision. Building on this line of work, PhoenixSR develops HDA and DRW for transfer into pre-existing, independently designed diffusion-free SR networks. Its contribution lies in adapting score estimation and guidance to this heterogeneous setting while preserving the student's original inference graph.

Difference from Cross-Space Distillation. Cross-Space Distillation (Nguyen et al., 2026) studies distribution-level knowledge transfer between diffusion teachers and one-step diffusion students whose latent resolutions or VAE spaces differ. Its heterogeneous setting therefore arises from incompatibility between different diffusion latent spaces, while the deployed student remains a diffusion-derived generator.

PhoenixSR considers a different form of heterogeneity: the student is not a diffusion model, but a native deterministic SR network such as SwinIR, HAT, Real-ESRGAN, or SeeMoRe. The diffusion latent space is used only as a supervision space during training and does not become part of the student architecture or inference process.

Taken together, PhoenixSR specifically studies *score-based distribution-level transfer of a pretrained diffusion prior into pre-existing, independently designed diffusion-free SR architectures while preserving their native feed-forward inference graphs*. HDA enables this distribution-level transfer to operate under the paired and fidelity-sensitive constraints of SR, while DRW further addresses sample-wise variation in the stability of online fake-score guidance.

Discussion of direct quantitative comparison. A strictly controlled comparison between heterogeneous distillation methods requires the same student initialization, training data, degradation pipeline, optimization budget, paired supervision, and evaluation protocol. This is particularly important in real-world SR, where changes in degradation synthesis and reconstruction objectives can substantially affect both distortion and perceptual metrics.

For DTKD (Park & Song, 2026), the published experiments use a diffusion SR teacher and a SwinIR-based student under their own training and distillation setting. At the time of submission, we could not identify an official public implementation or pretrained distillation checkpoint that would allow DTKD to be reproduced under the controlled PhoenixSR setting. Directly inserting its reported numbers into our quantitative tables would therefore mix different training protocols and would not constitute a controlled comparison. Instead, the main paper includes an output-level heterogeneous KD baseline trained with the same student initialization, training data, update budget, paired supervision, and model-selection protocol as PhoenixSR. This baseline is intended to isolate the difference between direct teacher-output supervision and distribution-level transfer, rather than to serve as a reproduction of DTKD.

Similarly, at the time of submission, we could not identify an official public implementation of DM-SR (Park & Lee, 2026) that could be evaluated under our unified protocol. Moreover, DM-SR and PhoenixSR target different deployment settings: DM-SR retains the pretrained diffusion model in its restoration pipeline, whereas PhoenixSR removes all diffusion-related components after training. We therefore include representative diffusion-based SR methods as inference-time reference baselines and use controlled SDS/DMD-based variants to analyze distribution-level transfer within the native DFSR setting.

These comparisons separate two related but distinct questions. Prior work has established that diffusion priors can improve perceptual SR and can be transferred to efficient students. PhoenixSR instead investigates whether a pretrained diffusion prior can be transferred through score-based distribution matching directly into independently designed native diffusion-free SR architectures, while preserving their original feed-forward inference path.

B Additional Method Details

B.1 Latent-Space Distribution Matching

PhoenixSR uses Stable Diffusion 2.1 as a training-only diffusion prior. Let the diffusion-free SR student produce

$$\hat{\mathbf{x}}_0 = G_\theta(\mathbf{y}), \tag{9}$$

and let E denote the frozen Stable Diffusion VAE encoder. The generated image is mapped to the latent space as

$$\hat{\mathbf{z}}_0 = E(\hat{\mathbf{x}}_0), \qquad \hat{\mathbf{z}}_t = \alpha_t \hat{\mathbf{z}}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{10}$$

The real-score model and fake-score model predict noise on the same noisy generated latent:

$$\boldsymbol{\epsilon}_{\text{real}} = \boldsymbol{\epsilon}_{\text{real}}(\hat{\mathbf{z}}_t, t, \mathbf{c}), \qquad \boldsymbol{\epsilon}_{\text{fake}} = \boldsymbol{\epsilon}_{\text{fake}}(\hat{\mathbf{z}}_t, t, \mathbf{c}), \tag{11}$$

where $\mathbf{c}$ is the text condition. In the reported experiments, all samples use the same fixed quality prompt.

Under noise prediction, the diffusion score is proportional to the negative predicted noise. Thus, the real–fake score discrepancy provides distribution-level supervision without requiring the SR student to share the diffusion architecture.

B.2 Denoised-Latent DMD Implementation

For stable optimization, our implementation realizes the score discrepancy through an equivalent denoised-latent parameterization. Given cumulative diffusion coefficient $\bar{\alpha}_t$, define

$$a_t = \sqrt{\bar{\alpha}_t}, \qquad b_t = \sqrt{1 - \bar{\alpha}_t}. \tag{12}$$

The real and fake predictions of the clean latent are

$$\tilde{\mathbf{z}}_0^{\text{real}} = \frac{\hat{\mathbf{z}}_t - b_t \boldsymbol{\epsilon}_{\text{real}}}{a_t}, \qquad \tilde{\mathbf{z}}_0^{\text{fake}} = \frac{\hat{\mathbf{z}}_t - b_t \boldsymbol{\epsilon}_{\text{fake}}}{a_t}. \tag{13}$$

We compute the per-sample distributional guidance

$$\mathbf{g}_i = \frac{\tilde{\mathbf{z}}_{0,i}^{\text{fake}} - \tilde{\mathbf{z}}_{0,i}^{\text{real}}}{\text{mean}\left(\left|\hat{\mathbf{z}}_{0,i} - \tilde{\mathbf{z}}_{0,i}^{\text{real}}\right|\right) + \epsilon}. \tag{14}$$

The denominator prevents low-noise samples with very small denoising distances from dominating optimization. For additional stability, the root-mean-square magnitude of $\mathbf{g}_i$ is clipped when it exceeds the configured threshold.

A detached surrogate target is then formed as

$$\mathbf{z}_{0,i}^{\text{tar}} = \text{sg}\left(\hat{\mathbf{z}}_{0,i} - \mathbf{g}_i\right), \tag{15}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. The corresponding per-sample DMD surrogate is

$$\mathcal{L}_{\text{DMD}}^i = \frac{1}{2}\left\|\hat{\mathbf{z}}_{0,i} - \mathbf{z}_{0,i}^{\text{tar}}\right\|_2^2. \tag{16}$$

Because the target is detached, differentiating Eq. (16) produces the desired distributional guidance without back-propagating through the score models.

B.3 Detailed Heterogeneous Distribution Adaptation

LoRA-adapted real-score model. The pretrained Stable Diffusion U-Net is used as the real-score model. Its original parameters remain frozen, and LoRA (Hu et al., 2022) adapters are inserted into the attention and projection layers. For an adapted linear layer,

$$\mathbf{W} = \mathbf{W}_0 + \frac{\gamma}{r}\mathbf{B}\mathbf{A}, \tag{17}$$

where $\mathbf{W}_0$ is frozen, $\mathbf{A}$ and $\mathbf{B}$ are trainable low-rank matrices, r is the rank, and γ is the LoRA scaling factor.

Given a paired HR image $\mathbf{x}$, its latent representation is

$$\mathbf{z}_0^{HR} = E(\mathbf{x}), \qquad \mathbf{z}_t^{HR} = \alpha_t \mathbf{z}_0^{HR} + \sigma_t \boldsymbol{\epsilon}. \tag{18}$$

Only the LoRA parameters are optimized using

$$\mathcal{L}_{\text{real}} = \mathbb{E}\left[\left\|\boldsymbol{\epsilon}_{\text{real}}(\mathbf{z}_t^{HR}, t, \mathbf{c}) - \boldsymbol{\epsilon}\right\|_2^2\right]. \tag{19}$$

This approach limits adaptation while retaining the pretrained backbone. As illustrated in Fig. 8(b), the frozen real-score model may transform restoration artifacts into perceptually plausible but unfaithful textures, whereas LoRA adaptation alleviates this behavior and provides guidance that is better aligned with SR reconstruction.

Representation-regularized fake-score model. The fake-score model is initialized by copying the pretrained diffusion U-Net and is fully optimized online to track the evolving student distribution:

$$\mathcal{L}_{\text{fake}} = \mathbb{E}\left[\left\|\boldsymbol{\epsilon}_{\text{fake}}(\hat{\mathbf{z}}_t, t, \mathbf{c}) - \boldsymbol{\epsilon}\right\|_2^2\right]. \tag{20}$$

Because the student distribution changes continuously, the fake-score model may otherwise overfit transient artifacts. We therefore introduce representation alignment following REPA (Yu et al., 2025). Let $\mathbf{f}^l$ denote an intermediate fake-UNet feature, $\mathcal{P}$ a trainable projection, and $\mathbf{h}_{\text{DINO}}$ the representation produced by a frozen DINOv3 encoder (Siméoni et al., 2026). The REPA objective is

$$\mathcal{L}_{\text{REPA}} = -\mathbb{E}\left[\frac{1}{N_p}\sum_{n=1}^{N_p} \text{sim}\left(\mathbf{u}_n^l, \text{sg}(\mathbf{v}_n)\right)\right], \tag{21}$$

where

$$\mathbf{u}^l = \mathcal{P}(\mathbf{f}^l), \qquad \mathbf{v} = \mathbf{h}_{\text{DINO}}(\hat{\mathbf{x}}_0), \tag{22}$$

and sim denotes cosine similarity. The complete fake-score objective is

$$\mathcal{L}_{\text{fake}}^{\text{total}} = \mathcal{L}_{\text{fake}} + \lambda_{\text{repa}}\mathcal{L}_{\text{REPA}}. \tag{23}$$

Sample-level fidelity anchor. Distribution matching alone does not guarantee correspondence with the paired HR target. We therefore constrain the SR generator using

$$\mathcal{L}_{\text{SFA}} = \lambda_{\text{pix}}\mathcal{L}_{\text{pix}} + \lambda_{\text{per}}\mathcal{L}_{\text{per}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}, \tag{24}$$

with

$$\mathcal{L}_{\text{pix}} = \mathbb{E}\left[\left\|G_\theta(\mathbf{y}) - \mathbf{x}\right\|_1\right], \tag{25}$$

and

$$\mathcal{L}_{\text{per}} = \mathbb{E}\left[\left\|\varphi(G_\theta(\mathbf{y})) - \varphi(\mathbf{x})\right\|_1\right], \tag{26}$$

where φ is the feature extractor used for perceptual supervision (Johnson et al., 2016). The adversarial component uses the same generator/discriminator objective as the underlying real-world SR training framework.

The student generator is therefore optimized jointly by sample-level fidelity supervision and distribution-level diffusion guidance:

$$\nabla_\theta \mathcal{L}_G = \lambda_{\text{dmd}}\nabla_\theta \mathcal{L}_{\text{DMD}} + \nabla_\theta \mathcal{L}_{\text{SFA}}. \tag{27}$$

B.4 Detailed Directional Reliability Weighting

DRW estimates whether the online fake-score model behaves consistently around the generated and real-data neighborhoods. For the i-th paired sample, let

$$\mathbf{z}_0^{G,i} = E(G_\theta(\mathbf{y}^i)), \qquad \mathbf{z}_0^{HR,i} = E(\mathbf{x}^i). \tag{28}$$



Figure 8: **Analysis of diffusion guidance for SR distillation.** (a) **Effect of timestep selection.** Low-noise states preserve more input-consistent structures, while stronger noise makes diffusion guidance increasingly dominated by the generative prior and may introduce unfaithful details. (b) **Visualization of real-score adaptation.** The frozen real-score model may convert restoration artifacts into plausible but unfaithful textures, while LoRA adaptation alleviates such artifact-to-texture degeneration and provides more SR-compatible guidance.

The same timestep and Gaussian noise are applied to both:

$$\mathbf{z}_t^{G,i} = \alpha_t \mathbf{z}_0^{G,i} + \sigma_t \boldsymbol{\epsilon}^i, \qquad \mathbf{z}_t^{HR,i} = \alpha_t \mathbf{z}_0^{HR,i} + \sigma_t \boldsymbol{\epsilon}^i. \tag{29}$$

Their fake-score denoising residuals are

$$\mathbf{r}_G^i = \boldsymbol{\epsilon}_{\text{fake}}(\mathbf{z}_t^{G,i}, t, \mathbf{c}^i) - \boldsymbol{\epsilon}^i, \tag{30}$$

and

$$\mathbf{r}_{HR}^i = \boldsymbol{\epsilon}_{\text{fake}}(\mathbf{z}_t^{HR,i}, t, \mathbf{c}^i) - \boldsymbol{\epsilon}^i. \tag{31}$$

We compute

$$s_i = \frac{\langle \mathbf{r}_G^i, \mathbf{r}_{HR}^i \rangle}{\|\mathbf{r}_G^i\|_2 \|\mathbf{r}_{HR}^i\|_2}. \tag{32}$$

Importantly, s_i is not an image-quality score. It measures the directional consistency of the fake-score model around generated and real-data samples. Higher consistency indicates that the fake-score model exhibits more similar local denoising behavior around the two regions.

The reliability weight is

$$\omega_i = \text{clip}\left(N \frac{\exp(s_i/\tau)}{\sum_{j=1}^N \exp(s_j/\tau)}, \omega_{\min}, \omega_{\max} \right). \tag{33}$$

The final weighted DMD loss is

$$\mathcal{L}_{\text{DMD}}^{\text{DRW}} = \frac{1}{N} \sum_{i=1}^N \text{sg}(\omega_i) \mathcal{L}_{\text{DMD}}^i. \tag{34}$$

During the generator update, no gradient is propagated through s_i or ω_i.

Distributed normalization. Training uses one image per GPU. Therefore, computing the softmax independently on each GPU would degenerate to $\omega_i = 1$. Instead, the directional similarities are gathered across distributed workers before the softmax. With eight GPUs and one sample per GPU, the normalization pool contains eight samples. Only the resulting scalar weights are used to modulate the per-sample DMD contributions.

DRW does not change the real–fake score direction of an individual sample. It only changes that sample's relative contribution to the aggregated update.

C Training Strategy and Implementation Details

C.1 Timestep Truncation and Noise Shift

The diffusion timestep controls the strength of the generative prior involved in distribution matching. For fidelity-sensitive SR, excessively noisy states may weaken the correspondence with the student reconstruction and make the guidance increasingly prior-driven. As shown in Fig. 8(a), low-noise states

preserve more input-consistent structures, whereas stronger noise leads to increasingly generative predictions and may introduce unfaithful details. We therefore restrict DMD supervision to

$$t \in [0, 200). \tag{35}$$

Within this range, the implementation applies a monotonic noise-shift transform. A base timestep t_b is sampled uniformly and normalized as

$$u = \frac{t_b - t_{\min}}{t_{\max} - t_{\min} - 1}. \tag{36}$$

The shifted normalized timestep is

$$\phi_s(u) = \frac{s\,u}{1 + (s-1)u}, \tag{37}$$

and the final discrete timestep is obtained by mapping $\phi_s(u)$ back to $[t_{\min}, t_{\max})$ and rounding to the nearest integer.

The shift parameter s is changed between training stages, allowing the distribution of sampled noise levels to vary during optimization while remaining within the fidelity-oriented range $[0, 200)$.

C.2 Warm-Up and Asynchronous Optimization

The SR generator and the two score models evolve at different rates. Applying DMD before the fake-score model has adapted to the current student distribution can produce unstable guidance. We therefore first warm up the score models while keeping generator updates disabled.

Following the two-time-scale training principle of DMD2 (Yin et al., 2024a), score models are refreshed more frequently than the SR generator after warm-up. The real-score LoRA is optimized on noisy HR latents, while the fake-score model is optimized on detached noisy student latents. The generator is updated only at its configured interval using the latest real/fake score estimates, DRW weights, and SFA supervision. The discriminator follows its own adversarial update schedule.

This asynchronous optimization reduces stale fake-score estimates and separates score-model fitting from generator optimization.

C.3 Complete Training Procedure

Algorithm 1 summarizes the complete PhoenixSR training procedure.

C.4 Experimental Configuration

The HR training set combines DIV2K (Agustsson & Timofte, 2017), Flickr2K (Lim et al., 2017), OST (Wang et al., 2018), WED (Ma et al., 2017), FFHQ (Karras et al., 2019), Manga109 (Matsui et al., 2017), and SCUT-CTW1500 (Liu et al., 2019). LR inputs are synthesized using the Real-ESRGAN degradation pipeline (Wang et al., 2021). Training crops are 256×256 in HR and 64×64 in LR for $4\times$ SR. Experiments are conducted on eight NVIDIA V100 GPUs with one sample per GPU.

Unless otherwise specified, we use the following default configuration. The SR generator is optimized with Adam (Kingma & Ba, 2015) using a learning rate of 3×10^{-5} and $(\beta_1, \beta_2) = (0.9, 0.99)$. The real-score U-Net uses LoRA with rank 16, scaling factor 32, and dropout 0, while the pretrained diffusion backbone remains frozen. The fake-score U-Net is initialized from the pretrained diffusion U-Net and is fully optimized online. Both score models are optimized with AdamW (Loshchilov & Hutter, 2019) using a learning rate of 3×10^{-5}.

For the sample-level fidelity anchor, we set $\lambda_{\text{pix}} = 0.1$, $\lambda_{\text{per}} = 0.01$, and $\lambda_{\text{adv}} = 0.1$. The distribution-matching weight is set to $\lambda_{\text{dmd}} = 1.0$, and the REPA regularization weight is $\lambda_{\text{repa}} = 0.1$. For DRW, the temperature is set to $\tau = 0.2$, and the sample weights are clipped to $[0.2, 3.0]$. DMD supervision uses the truncated timestep range $[0, 200)$ together with the noise-shift strategy described above.

The reported experiments use the same fixed text prompt for all samples:

Algorithm 1 PhoenixSR Distillation

Require: Paired data $\mathcal{D}$; DFSR student G_θ; VAE encoder E; real-score model; fake-score model; DINOv3 encoder; discriminator.

Initialize G_θ from the corresponding DFSR checkpoint.
Initialize the real-score and fake-score models from Stable Diffusion 2.1.
Freeze the real-score backbone and optimize only its LoRA adapters.
Fully optimize the fake-score model with denoising and REPA objectives.
for training iteration k **do**
  Sample $(\mathbf{y}, \mathbf{x}) \sim \mathcal{D}$.
  Compute $\hat{\mathbf{x}}_0 = G_\theta(\mathbf{y})$.
  Encode $\hat{\mathbf{z}}_0 = E(\hat{\mathbf{x}}_0)$ and $\mathbf{z}_0^{HR} = E(\mathbf{x})$.
  Sample $t \in [0, 200)$ and apply the noise-shift transform.
  Update the real-score LoRA using noisy HR latents.
  Update the fake-score model using detached noisy student latents and REPA.
  if generator update is enabled after warm-up **then**
    Evaluate real/fake predictions on the same noisy generated latent.
    Compute per-sample DMD guidance using Eq. (14).
    Compute DRW weights using Eqs. (32)–(33).
    Compute the sample-level fidelity anchor in Eq. (24).
    Update G_θ with weighted DMD and SFA supervision.
  end if
  Update the discriminator according to the adversarial schedule.
end for
Discard VAE, diffusion score models, LoRA adapters, DINO/REPA modules, and discriminator.
Deploy G_θ only.

> "masterpiece, best quality, ultra-detailed, high resolution, highly detailed, sharp focus, extremely intricate, professional photography, high-definition."

Training cost. The training cost varies across student backbones because of their different model capacities and convergence behaviors. In our experiments, PhoenixSR distillation typically requires approximately 3–6 days on eight NVIDIA V100 GPUs, with training lengths ranging from roughly 10^5 to 10^6 iterations. The additional cost is confined to the offline distillation stage, during which the diffusion score models, VAE encoder, DINO/REPA modules, and discriminator are used only to provide training supervision. After distillation, all auxiliary modules are discarded, and only the original feed-forward SR student is retained for inference. Therefore, PhoenixSR introduces no additional model parameters, FLOPs, or inference latency at deployment.

Further implementation and evaluation details. For finer-grained implementation and evaluation details, including more hyperparameter details, training schedules, model-selection settings, metric configurations, and efficiency measurement, we refer to the source code and configuration files. These materials can be provided to reviewers and area chairs through an anonymous repository upon request during the discussion phase.

D Additional Ablation Studies

Unless otherwise specified, the ablation experiments use SwinIR-L under the same controlled 80K-iteration setting as the main component ablation.

D.1 Robustness Across Random Seeds

Since distribution-matching objectives involve stochastic diffusion perturbations and online score-model optimization, we additionally evaluate the robustness of PhoenixSR across independent training runs. We repeat the distillation experiments with three random seeds for one representative large student (SwinIR-L) and one lightweight student (SeeMoRe) on RealSR. The pretrained base checkpoints are deterministic references and are therefore evaluated once, while the distillation

Table 7: **Robustness across random seeds on RealSR.** PhoenixSR variants are independently trained with three random seeds and reported as mean±standard deviation. The base checkpoints are fixed references evaluated once.

Student	Method	PSNR↑	LPIPS↓	MANIQA↑	MUSIQ↑	CLIP-IQA↑
SwinIR-L	Base checkpoint	26.30	0.253	0.522	58.6	0.45
	PhoenixSR w/o DRW	26.20±0.03	0.251±0.004	0.557±0.007	62.9±0.3	0.481±0.005
	PhoenixSR	26.40±0.04	0.247±0.008	0.568±0.006	64.1±0.5	0.493±0.007
SeeMoRe	Base checkpoint	26.0	0.295	0.347	48.7	0.381
	PhoenixSR w/o DRW	25.41±0.07	0.277±0.009	0.511±0.008	60.9±0.4	0.519±0.009
	PhoenixSR	25.94±0.03	0.269±0.005	0.531±0.005	61.7±0.3	0.530±0.006

Table 8: **DRW versus uniform weighting and hard sample selection.**

Method	PSNR↑	MANIQA↑	MUSIQ↑	CLIP-IQA↑
SFA only	26.3124	0.5067	56.8245	0.4538
Uniform DMD w/o DRW	26.1587	0.5678	63.5729	0.4963
Low-rel. DMD only	26.2072	0.5881	64.9234	0.5173
High-rel. DMD only	26.3973	0.5526	61.9648	0.4892
Full PhoenixSR w/ DRW	26.4409	0.5745	64.3102	0.5026

Table 9: **Relationship between the DRW score and DMD guidance stability.** Samples are divided into five equal-sized groups according to the directional consistency score s_i.

DRW-score group	Mean s_i	Guidance stability ↑
Lowest 20%	-0.604	0.074
20–40%	-0.195	0.091
40–60%	0.206	0.173
60–80%	0.579	0.418
Highest 20%	0.773	0.577

variants are independently trained for each seed. We report the mean and standard deviation over the three runs.

Table 7 shows that the improvements of PhoenixSR are consistent across independent runs, with small run-to-run variation relative to the observed gains. In particular, the comparison between PhoenixSR with and without DRW verifies that the contribution of reliability weighting is not tied to a single random seed. These results complement the single-run results in the main paper and provide additional evidence that the reported gains are robust to stochastic training variation.

D.2 Soft Weighting Versus Hard Selection

We compare DRW with uniform DMD weighting and hard reliability-based sample selection. For hard selection, samples are split according to the same residual-direction consistency used by DRW.

Uniform DMD improves perceptual metrics but reduces PSNR relative to SFA alone. High-reliability-only training preserves fidelity better but provides weaker perceptual improvement, whereas low-reliability-only training moves further toward the perceptual regime. DRW retains contributions from both subsets and softly adjusts their relative strength, obtaining the highest PSNR together with strong perceptual quality. This also shows that the directional consistency should be interpreted as a reliability signal for score estimation rather than a direct image-quality measure.

D.3 Empirical Validation of DRW Reliability

The experiments above show that DRW improves the final SR performance. We further examine whether the directional consistency score in Eq. (32) indeed reflects the reliability of the sample-wise DMD guidance. Recall that the per-sample DMD guidance $\mathbf{g}_i$ is defined in Eq. (14). Intuitively, if the fake-score model provides reliable local guidance for a sample, independently sampled diffusion perturbations should produce more consistent DMD update directions.

Directional consistency versus DMD guidance stability. For each validation sample, we first compute its DRW score s_i using one sampled timestep and Gaussian noise. We then independently sample $K = 8$ additional perturbations and compute the corresponding DMD guidance vectors $\{\mathbf{g}_i^{(1)}, \ldots, \mathbf{g}_i^{(K)}\}$. We measure the guidance stability of sample i by the average pairwise cosine

Table 10: **Comparison of sample-wise reliability proxies.** We report the Spearman rank correlation between each proxy and DMD guidance stability. A larger positive correlation indicates that the proxy better reflects guidance stability.

Reliability proxy	Spearman $\rho\uparrow$
Latent distance $-\|\mathbf{z}_0^G - \mathbf{z}_0^{HR}\|_2$	0.33
Residual distance $-\|\mathbf{r}_G - \mathbf{r}_{HR}\|_2$	0.27
Prediction cosine $\cos(\boldsymbol{\epsilon}_{\text{fake}}(\mathbf{z}_t^G), \boldsymbol{\epsilon}_{\text{fake}}(\mathbf{z}_t^{HR}))$	0.41
Residual-direction cosine (DRW) $\cos(\mathbf{r}_G, \mathbf{r}_{HR})$	**0.59**

Table 11: **Comparison of HDA score-model designs.** We compare different configurations of the real and fake score models, including full fine-tuning, REPA adaptation, LoRA adaptation, and SFA.

Method	PSNR↑	MANIQA↑	MUSIQ↑	CLIP-IQA↑
Frozen Real + Full Fake w/o SFA	21.2561	0.4023	39.6617	0.3569
Frozen Real + Full Fake + SFA	25.9778	0.5422	63.9735	0.4931
Frozen Real + REPA Fake + SFA	26.3012	0.5787	64.1157	0.4898
LoRA Real + REPA Fake + SFA	26.4409	0.5745	64.3102	0.5026

similarity:

$$\text{Stab}_i = \frac{2}{K(K-1)} \sum_{1 \le a < b \le K} \cos\left(\mathbf{g}_i^{(a)}, \mathbf{g}_i^{(b)}\right). \tag{38}$$

A larger value indicates that the DMD guidance remains more consistent under different diffusion perturbations.

We sort samples according to s_i and divide them into five equally sized groups. Table 9 reports the average guidance stability in each group.

As shown in Table 9, samples with higher directional consistency exhibit increasingly stable DMD guidance across stochastic perturbations. This observation supports our use of s_i as an empirical reliability proxy: a high DRW score indicates that the online fake-score model produces more consistent local guidance around the generated and paired HR regions. Importantly, this analysis does not interpret s_i as an image-quality measure; instead, it directly relates s_i to the stability of the distributional supervision used to optimize the student.

Comparison with alternative reliability proxies. We further examine whether the proposed residual-direction consistency is more informative than simpler alternatives. We compare four sample-wise signals: latent distance between the generated and HR latents, residual distance between $\mathbf{r}_G$ and $\mathbf{r}_{HR}$, cosine similarity between the raw fake-score predictions, and the proposed cosine similarity between denoising residuals. For distance-based measures, we use the negative distance such that a larger value consistently represents higher estimated reliability.

For each proxy, we compute its Spearman rank correlation with the guidance stability defined in Eq. (38). The results are reported in Table 10.

Table 10 shows that the proposed residual-direction consistency has the strongest correlation with DMD guidance stability among the tested signals. Compared with latent or residual distances, directional consistency focuses on whether the fake-score model provides compatible denoising directions around the generated and HR neighborhoods. Removing the shared noise component through the residual formulation also makes this comparison less dominated by the sampled perturbation itself. These results provide empirical support for using residual-direction consistency, rather than a generic sample-distance measure, to construct the DRW weights.

D.4 HDA Score-Model Designs

Table 11 compares the score-model configurations used during development of HDA.

DRW is kept enabled for all variants in this comparison so that only the score-model designs within HDA are varied. Without SFA, direct distribution matching produces severe degradation, showing that paired reconstruction constraints are critical for fidelity-sensitive SR. Adding SFA restores stable reconstruction and substantially improves perceptual quality. Representation regularization further improves PSNR and MANIQA, while the final LoRA-adapted real-score configuration achieves the strongest PSNR and CLIP-IQA with competitive perceptual scores. Together, these results support the use of fidelity anchoring, representation-regularized fake-score tracking, and lightweight real-score adaptation.

Table 12: **Ablation of the training schedule.**

Method	PSNR↑	MANIQA↑	MUSIQ↑	CLIP-IQA↑
w/o timestep truncation	25.6564	0.5439	61.6731	0.4733
w/o noise-shift curriculum	26.3797	0.5503	62.7315	0.4977
w/o warm-up	26.4011	0.5637	64.1149	0.4893
Full schedule	26.4409	0.5745	64.3102	0.5026

Table 13: **Comparison with conventional distillation strategies.**

Method	PSNR↑	MANIQA↑	MUSIQ↑	CLIP-IQA↑
SFA only	26.3124	0.5067	56.8245	0.4538
Feature KD	26.1638	0.5132	53.7781	0.4569
Output KD	25.9754	0.4983	56.7369	0.4691
DMD-only	21.2561	0.4023	39.6617	0.3569
PhoenixSR	26.4409	0.5745	64.3102	0.5026

D.5 Effect of the Training Schedule

Removing timestep truncation causes the largest PSNR degradation, supporting the use of low-noise diffusion states for fidelity-sensitive SR. Removing either the noise-shift strategy or warm-up produces smaller but consistent performance drops. The full schedule gives the strongest overall result among the tested variants.

D.6 Comparison with Conventional Distillation

We additionally compare PhoenixSR with conventional feature- and output-level knowledge distillation. Feature KD matches intermediate diffusion and DFSR representations using lightweight adapters, while output KD directly supervises the DFSR student with sampled diffusion-SR outputs.

Scope of the distribution-matching baselines. We use *DMD-only* to denote the unanchored score-difference ablation without paired SR fidelity losses or teacher-output regression, referred to as *vanilla DMD* in the main text and Fig. 3. This ablation isolates the role of fidelity anchoring in our SR setting; it is not a reproduction of the complete DMD or DMD2 training recipe (Yin et al., 2024b;a). The *DMD2 + SFA* baseline in Table 4 is a DMD2-based adaptation to native DFSR students with paired SR supervision. Its results concern this controlled transfer setting rather than the original diffusion-derived generators.

Feature KD provides only limited and inconsistent gains, reflecting the weak correspondence between diffusion and feed-forward SR representations. Output KD slightly improves CLIP-IQA but degrades PSNR, MANIQA, and MUSIQ relative to SFA, indicating that direct imitation of sampled diffusion outputs does not provide a consistently favorable fidelity–perception trade-off. DMD-only shows severe fidelity degradation in this unanchored SR setting. PhoenixSR achieves the strongest overall balance, supporting distribution-level transfer with fidelity and reliability control.

E Additional Results

E.1 Additional Quantitative Discussion

RealSR. For medium and large students, PhoenixSR preserves or improves PSNR while consistently strengthening perceptual quality. SwinIR-L, HAT, Real-ESRGAN, and SwinIR-M improve PSNR by 0.44, 0.22, 0.26, and 0.45 dB, respectively. Compact SwinIR-S and SeeMoRe remain within 0.2 dB of their original PSNR while obtaining substantially stronger no-reference quality scores.

DRealSR. SwinIR-L and SwinIR-M gain 0.65 and 0.54 dB in PSNR, while all medium/large backbones improve across the reported no-reference metrics. The perceptual gain is especially strong for compact students: SeeMoRe improves MUSIQ by 16.27, LIQE-Mix by 1.78, ARNIQA by 0.255, and TOPIQ by 0.188.

DIV2K-Val. PhoenixSR produces a stronger distortion–perception trade-off on DIV2K-Val. SwinIR-L and SwinIR-M improve both PSNR and perceptual quality. SwinIR-S and SeeMoRe exhibit PSNR changes of only -0.14 and -0.19 dB while improving MUSIQ by 17.73 and 23.33, respectively. SeeMoRe also improves LPIPS by 0.179.

These results further indicate that diffusion-prior transfer is particularly beneficial for compact models whose limited capacity otherwise restricts realistic high-frequency reconstruction.

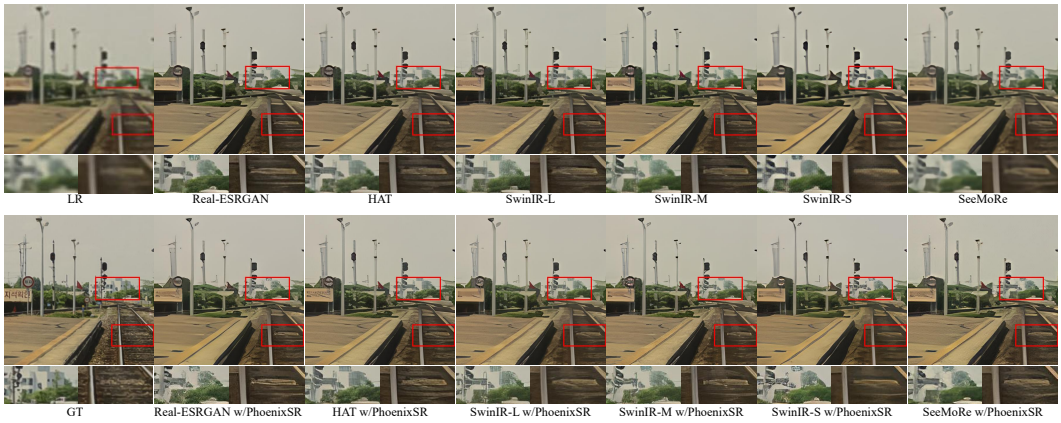


Figure 9: **Additional qualitative comparisons on RealSR.** PhoenixSR improves fine structures and local texture reconstruction across different feed-forward SR backbones while preserving the underlying scene content and geometry. Best viewed by zooming in.

E.2 Additional Qualitative Comparisons

We provide additional qualitative comparisons on RealSR and DRealSR to complement the quantitative results in the main paper. The examples cover different scene contents and local structures, and compare the original DFSR backbones with their PhoenixSR-distilled counterparts. Across different student architectures, PhoenixSR generally recovers clearer high-frequency details and more natural textures while preserving the underlying image structures.

RealSR. Figure 9 presents additional qualitative comparisons on the RealSR dataset. Compared with the original feed-forward SR models, PhoenixSR improves the restoration of fine structures and local textures across different student backbones. In particular, distant structures and rail-like textures become more distinguishable after distillation, while the overall geometry and scene content remain consistent with the original reconstruction. These visual improvements are consistent with the quantitative results reported on RealSR.

DRealSR. Figure 10 shows additional qualitative results on DRealSR. PhoenixSR produces clearer local structures and more recognizable texture details, particularly in regions that are easily over-smoothed by the original DFSR models. The improvements are observed across different student architectures, further demonstrating that the proposed diffusion-prior transfer is not tied to a specific feed-forward SR backbone.

Overall, these qualitative results are consistent with the quantitative observations across architectures. Improvements can be observed for CNN, Transformer, and lightweight students, providing additional evidence that PhoenixSR can transfer useful diffusion priors to different native feed-forward SR architectures without modifying their inference pipelines.

F User Study Details

Study protocol. We conduct a subjective study with 20 adult participants using 20 representative source images, including 7 images from RealSR, 7 from DRealSR, and 6 from DIV2K-Val. For each source image and DFSR backbone, the outputs before and after PhoenixSR distillation are presented in randomized left–right order to reduce positional bias. Participants are not informed which output is produced by PhoenixSR.

Participants score the overall visual quality of each result on a scale from 0 to 10, considering texture naturalness, detail sharpness, and artifact suppression. Each participant evaluates 120 paired comparisons, corresponding to 20 source images and six DFSR backbones. This results in 2,400 paired evaluations and 4,800 individual output ratings in total. All participants participate voluntarily

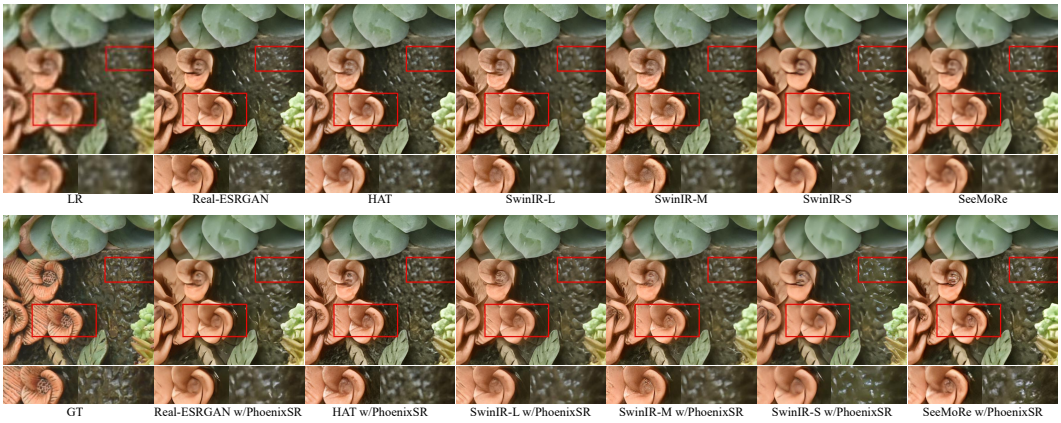


Figure 10: **Additional qualitative comparisons on DRealSR.** PhoenixSR recovers clearer local structures and more natural texture details across different feed-forward SR backbones while largely preserving reconstruction fidelity. Best viewed by zooming in.

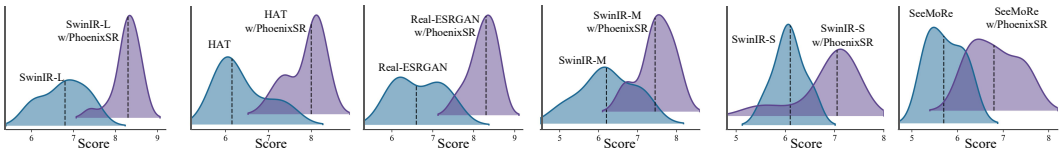


Figure 11: **User study** across six DFSR backbones.

after being informed of the purpose and procedure of the study, and no personally identifiable information is collected.

Image-level statistical analysis. For each source image and backbone, we first average the subjective scores from the 20 participants, yielding one image-level score for the original model and one for PhoenixSR. We report the mean score across the 20 source images together with a 95% confidence interval estimated by bootstrap resampling over images with 10,000 repetitions. Statistical significance is evaluated using a two-sided Wilcoxon signed-rank test on the paired image-level scores.

Table 14: **Detailed user-study statistics.** Scores are reported as image-level mean scores with 95% bootstrap confidence intervals. Δ denotes PhoenixSR minus the original backbone.

Backbone	Original	PhoenixSR	Δ (95% CI)
Real-ESRGAN	6.38 [6.14, 6.64]	7.89 [7.71, 8.06]	+1.51 [1.36, 1.66]
HAT	6.78 [6.56, 7.00]	8.29 [8.15, 8.40]	+1.51 [1.39, 1.63]
SwinIR-L	6.67 [6.43, 6.92]	8.24 [8.10, 8.37]	+1.57 [1.44, 1.70]
SwinIR-M	6.22 [5.92, 6.51]	7.50 [7.31, 7.67]	+1.28 [1.14, 1.42]
SwinIR-S	6.08 [5.95, 6.21]	6.87 [6.60, 7.11]	+0.79 [0.63, 0.92]
SeeMoRe	5.74 [5.58, 5.89]	6.81 [6.60, 7.04]	+1.07 [0.87, 1.31]
Overall	6.31 [6.10, 6.52]	7.60 [7.43, 7.75]	+1.29 [1.19, 1.38]

Across the 20 source images, PhoenixSR obtains a higher overall subjective score than the corresponding original DFSR outputs (7.60 versus 6.31), with a mean paired improvement of 1.29 points (95% CI: [1.19, 1.38]). The paired image-level difference is statistically significant under a two-sided

Wilcoxon signed-rank test ($p < 0.001$). Figure 11 and Table 14 further show that the subjective improvement is consistently observed across all six evaluated backbones.

Participant-level robustness analysis. We additionally average each participant's scores across all evaluated source images and backbones, yielding one paired Original/PhoenixSR score for each of the 20 participants. At the participant level, the mean subjective score increases from 6.31 for the original DFSR outputs to 7.60 for PhoenixSR, corresponding to an average improvement of 1.29 points. Bootstrap resampling over participants with 10,000 repetitions gives a 95% confidence interval of [1.05, 1.47] for the paired improvement. The difference remains statistically significant under a two-sided Wilcoxon signed-rank test ($p < 0.001$), and 18 of the 20 participants assign a higher average score to PhoenixSR. This complementary analysis indicates that the observed preference is consistent across participants rather than being driven only by a small subset of source images.

The subjective study is used only as complementary evidence of perceptual quality and does not replace the objective fidelity and perceptual evaluations reported in the main paper.