Disentangled Textual Priors for Diffusion-based Image Super-Resolution

1. More Implementation Details

Our DTPSR framework is implemented using PyTorch with mixed-precision training (fp16) to reduce memory usage and accelerate computation. The random seed is fixed to 6666 for reproducibility. We use the Adam optimizer with β_1=0.9, β_2=0.999, weight decay of 1×10^{-2}, and ϵ=1×10^{-8}. The conditioning scale for textual guidance is set to 1.0 by default, which provides a good balance between semantic alignment and generation quality.

2. Details of Multi-Branch Negative Prompts

To support the proposed Multi-Branch Classifier-Free Guidance (CFG) strategy described in the main paper, we define three category-specific negative prompts corresponding to the global, low-frequency, and high-frequency branches. These prompts are used during sampling to suppress undesired artifacts and improve semantic alignment at different abstraction levels.

Specifically, we use the following configurations:

- **Global Negative Prompt:** blurry, dotted, noise, raster lines, unclear, lowres, over-smoothed
- **Low-Frequency Negative Prompt:** blurry, misshaped object, bad anatomy, inconsistent lighting, unnatural shading, distorted global structure, low frequency banding, uneven illumination
- **High-Frequency Negative Prompt:** fake texture, excessive details, harsh edges, ringing artifacts, noise, aliasing, sharpening artifacts, hallucinated texture, jaggies

Each prompt is carefully crafted to suppress generation artifacts specific to its corresponding semantic domain. The global prompt mitigates overall blur and resolution issues; the low-frequency prompt focuses on structural distortions and lighting inconsistencies; and the high-frequency prompt targets unnatural textures and aliasing. This branch-specific suppression allows for more controllable and disentangled restoration, improving both semantic fidelity and perceptual quality without additional training overhead.

3. More Quantitative Comparisons

We further evaluate our method on the RealLR200 dataset to assess its generalization ability under challenging real-world degradations. As shown in Table 1, our DTPSR achieves the best performance across all three perceptual metrics, including MUSIQ, MANIQA, and CLIP-IQA. In particular, DTPSR surpasses the second-best method by a clear margin in MUSIQ (+2.03 over SeeSR), indicating superior aesthetic quality. For MANIQA and CLIP-IQA, DTPSR outperforms the closest competitor by +0.055 and +0.047, respectively, reflecting its ability to recover high-fidelity details while maintaining semantic alignment. These results demonstrate that the proposed disentangled textual priors and multi-branch guidance strategy generalize well to unseen real-world scenarios, yielding perceptually superior reconstructions even without dataset-specific tuning.

4. More Qualitative Comparisons

To further demonstrate the effectiveness of our DTPSR framework, we provide additional qualitative comparisons with state-of-the-art methods under various challenging real-world degradation scenarios. Figures 1 and 2 showcase more examples beyond those presented in the main paper. Our method consistently reconstructs sharper edges, more faithful textures, and fewer artifacts compared to competing approaches, particularly in regions with complex structures or high-frequency details.

5. Additional Ablation Studies

5.1. Extended Results of Main Paper Ablations

In the main paper, due to space limitations, we only presented the quantitative ablation results on the DRealSR dataset. In this section, we provide additional quantitative comparisons on the RealSR dataset to further validate our findings. Moreover, we include visual comparisons for each ablation setting, offering a more comprehensive understanding of their qualitative effects.

5.1.1. Effect of Global-Local Textual Priors

We further investigate the impact of incorporating global and local textual priors using the RealSR dataset. As shown in Table 2, removing both priors (Exp 2-1) yields the lowest perceptual quality, indicating that semantic guidance is crucial for real-world SR. Introducing only local priors (Exp 2-2) significantly boosts MANIQA (+0.0476) and CLIP-IQA (+0.0440) compared to Exp 2-1, demonstrating

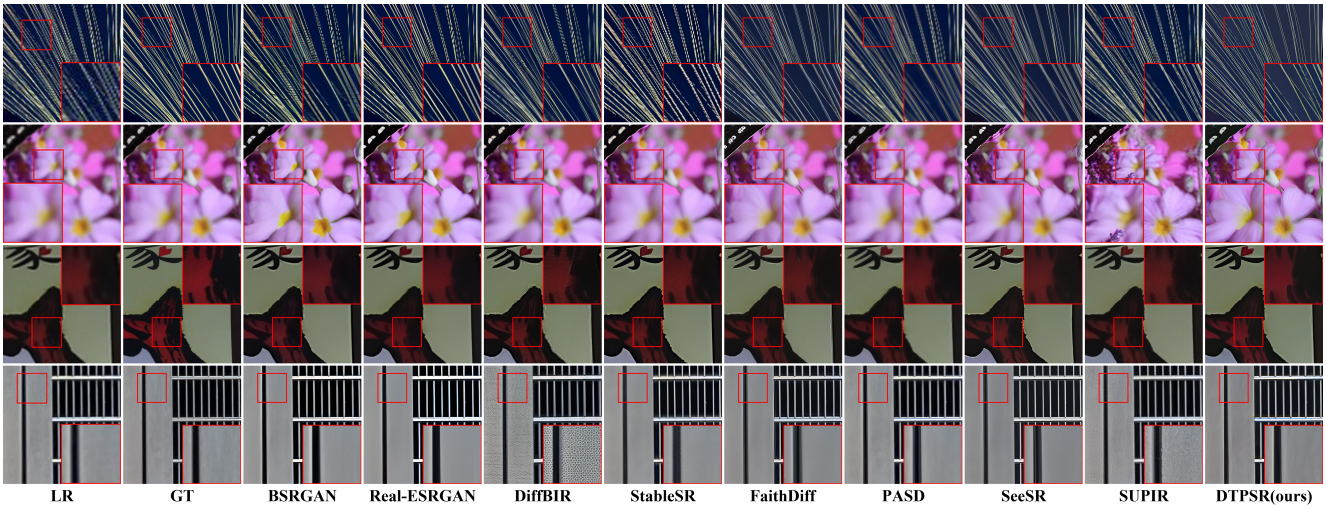


Figure 1. Additional qualitative comparisons with competing methods on challenging cases. DTPSR better preserves structural integrity and recovers realistic textures.

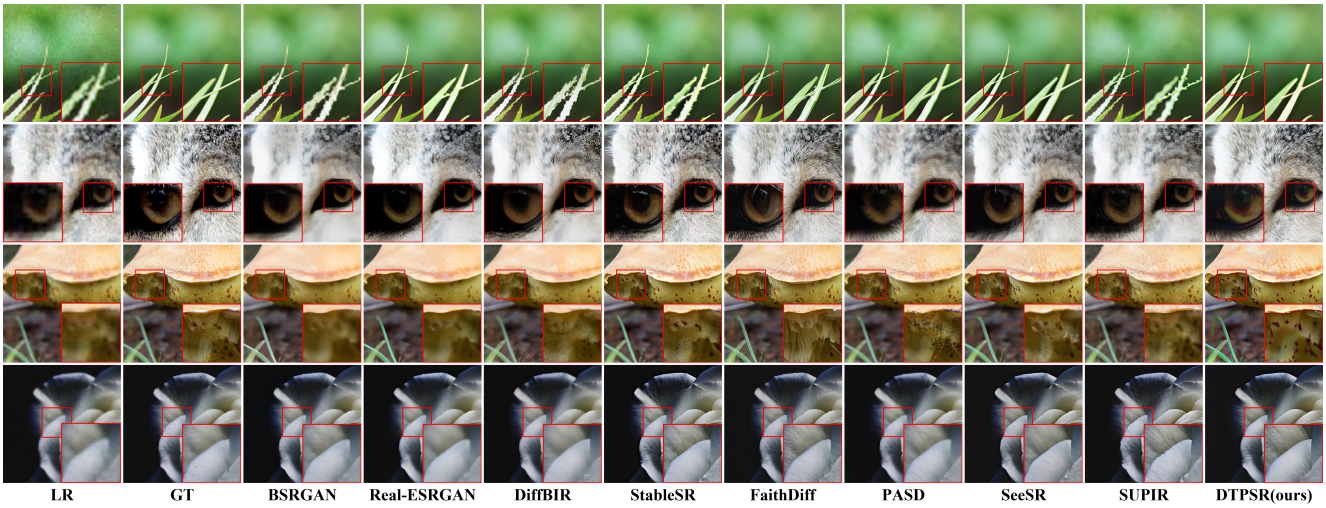


Figure 2. Additional qualitative comparisons with competing methods. DTPSR suppresses artifacts while producing perceptually pleasing details.

their effectiveness in refining object-level details. In contrast, using only global priors (Exp 2-3) results in limited improvements, suggesting that layout-level semantics alone are insufficient for restoring fine textures. The best performance is achieved when both global and local priors are used (Exp 2-4), with consistent gains across all perceptual metrics, confirming their complementary nature.

Qualitative comparisons in Fig. 3 illustrate this complementarity. Without global priors, outputs may exhibit locally accurate textures but inconsistent overall structure. Without local priors, restored images may preserve correct global layout but lack detailed textures. Combining both priors enables faithful reconstruction of global arrangement while maintaining realistic, high-quality local details. For example, in the first row of Fig. 3, when global information is absent, the local details of the plant branches appear clear but fail to align well with the ground truth. When local information is missing, the branches lose their fine details and become blurry. In the second row, without global information, the model misinterprets the stamen as petal textures, while without local information, it generates many hallucinated stamens. Only our approach, which integrates both global and local information, achieves the most faithful and visually pleasing super-resolution results.

5.1.2. Effect of Frequency-Aware Textual Priors

We further examine the contribution of low-frequency (LF) and high-frequency (HF) textual priors on the RealSR

Dataset	Metric	Real-ESRGAN	BSRGAN	StableSR	DiffBIR	SeeSR	SUPIR	FaithDiff	Ours
RealLR200	MUSIQ↑	61.40	64.86	51.61	66.43	<u>69.55</u>	62.29	69.06	**71.58**
	MANIQA↑	0.3540	0.3704	0.2955	<u>0.5037</u>	0.4986	0.4193	0.4281	**0.5591**
	CLIP-IQA↑	0.4832	0.5698	0.4394	<u>0.6961</u>	0.6830	0.6030	0.6617	**0.7435**

Table 1. Quantitative comparison on the RealLR200 dataset. The best and second-best results for each metric are highlighted in **bold** and <u>underlined</u>, respectively.

Figure 3. Visual comparison of different combinations of global and local textual priors. Using both enables coherent global structure and realistic local details.

Exp	Global	Local	MANIQA↑	CLIP-IQA↑	MUSIQ↑
2-1	×	×	0.5417	0.6834	70.45
2-2	×	✓	0.5893	0.7274	71.52
2-3	✓	×	0.5518	0.6827	70.70
2-4	✓	✓	**0.6021**	**0.7278**	**71.84**

Table 2. Ablation study on the effect of global and local textual priors on the RealSR dataset.

Exp	HF	LF	MANIQA↑	CLIP-IQA↑	MUSIQ↑
3-1	×	✓	0.5864	0.7161	71.69
3-2	✓	×	0.5650	0.6961	71.08
3-3	✓	✓	**0.6021**	**0.7278**	**71.84**

Table 3. Ablation study on the effect of frequency-aware textual priors on the RealSR dataset.

dataset. As shown in Table 3, removing HF priors (Exp 3-1) results in notable drops in MANIQA and CLIP-IQA, indicating degraded texture quality. Without HF guidance, the model struggles to produce correct fine textures. For

Exp	Type	MANIQA↑	CLIP-IQA↑	MUSIQ↑
4-1	Mixed	0.6020	0.7244	71.76
4-2	Disentangled	**0.6021**	**0.7278**	**71.84**

Table 4. Comparison of frequency-mixed and frequency-disentangled learning on the RealSR dataset.

Exp	CFG Strategy	MANIQA↑	CLIP-IQA↑	MUSIQ↑
5-1	None	0.6163	0.6890	71.53
5-2	Single	0.5922	0.7223	71.54
5-3	Multi (Ours)	**0.6021**	**0.7278**	**71.84**

Table 5. Ablation study on multi-branch classifier-free guidance (CFG) on the RealSR dataset.

instance, in Fig. 4, the first-row chair exhibits inaccurate metallic patterns, and in the second row, erroneous textures appear on the wall. This highlights the importance of HF priors for realistic detail synthesis.

Conversely, removing LF priors (Exp 3-2) leads to over-production of non-existent details, reflected by increased

Order	RealSR			DRealSR		
	MANIQA↑	CLIP-IQA↑	MUSIQ↑	MANIQA↑	CLIP-IQA↑	MUSIQ↑
Local → Global	**0.6242**	0.7210	70.5662	0.5986	0.7335	68.5563
Global → Local (Ours)	0.6021	**0.7278**	**71.8387**	**0.6011**	**0.7640**	**69.2433**

Table 6. Ablation study on the injection order of global and local textual priors.

Order	RealSR			DRealSR		
	MANIQA↑	CLIP-IQA↑	MUSIQ↑	MANIQA↑	CLIP-IQA↑	MUSIQ↑
HF→LF	0.5860	0.7072	71.7024	0.5701	0.7462	68.2454
LF→HF (ours)	**0.6021**	**0.7278**	**71.8387**	**0.6011**	**0.7640**	**69.2433**

Table 7. Ablation study on the effect of low–high frequency prior injection order.

artifacts in both rows of Fig. 4. In the first row, spurious textures emerge on smooth regions of the chair, and in the second row, the wall surface becomes cluttered with false patterns. Such over-generation confirms that LF priors are essential for constraining structure and suppressing hallucinations.

Our full model (Exp 3-3), which integrates both HF and LF priors, achieves the highest scores across all perceptual metrics—0.6021 MANIQA, 0.7278 CLIP-IQA, and 71.84 MUSIQ—and delivers the most visually faithful results. The combined frequency-aware priors allow the model to preserve coherent structures while generating accurate, high-quality textures, striking the right balance between detail fidelity and artifact suppression.

5.1.3. Disentangled vs. Mixed Frequency-Aware Learning

We further compare our disentangled frequency-aware learning strategy with a frequency-mixed counterpart on the RealSR dataset. As shown in Table 4, the disentangled design achieves slightly higher scores across all perceptual metrics—0.6021 MANIQA, 0.7278 CLIP-IQA, and 71.84 MUSIQ—compared to the mixed setting. Although the numerical gains may appear marginal, the improvement is consistent, indicating that explicitly separating low- and high-frequency priors yields more stable perceptual benefits.

Qualitative comparisons in Fig. 5 highlight clearer advantages. In the first row, the disentangled model successfully reconstructs the correct window textures, while the mixed model produces oversimplified and less realistic patterns. In the second row, the mixed approach introduces hallucinated cactus spines that do not exist in the ground truth, whereas the disentangled model avoids such artifacts and maintains clean, realistic object surfaces. These examples demonstrate that modeling low- and high-frequency semantics separately not only improves structural fidelity but also leads to more accurate and artifact-free texture synthesis.

5.1.4. Effect of Multi-Branch Negative Prompts

We further evaluate the impact of our proposed multi-branch classifier-free guidance (CFG) strategy on the RealSR dataset. As shown in Table 5, removing negative prompts entirely (Exp 5-1) yields suboptimal perceptual quality due to the lack of semantic suppression, leading to visible inconsistencies in both global layout and local details. Introducing a single generic negative prompt (Exp 5-2) improves CLIP-IQA, indicating enhanced semantic alignment, but sacrifices MANIQA and fails to fully suppress frequency-specific artifacts.

Our full approach (Exp 5-3) adopts three distinct negative prompts targeting global, low-frequency, and high-frequency branches, respectively. This disentangled suppression mechanism achieves the highest scores across all three perceptual metrics—0.6021 MANIQA, 0.7278 CLIP-IQA, and 71.84 MUSIQ—demonstrating its effectiveness in balancing semantic consistency with perceptual fidelity.

The qualitative comparisons in Fig. 6 further support these findings. Without negative prompts, the outputs tend to exhibit structural distortions and noisy textures. Single-branch CFG reduces some artifacts but may oversmooth fine details. In contrast, our multi-branch CFG preserves realistic textures while maintaining coherent global structure, validating its advantage in frequency-aware, disentangled guidance.

5.2. New Ablation Studies

5.2.1. Effect of Global-Local Prior Injection Order

We further investigate the impact of the injection order of global and local textual priors on super-resolution performance. Two strategies are compared: (1) *Local → Global*, where local priors are injected before global priors, and (2) *Global → Local (Ours)*, where global priors are injected first to establish the overall structural layout, followed by local priors for refining fine-grained details.

Quantitative results in Table 6 show that the *Global → Local* ordering outperforms the *Local → Global* variant

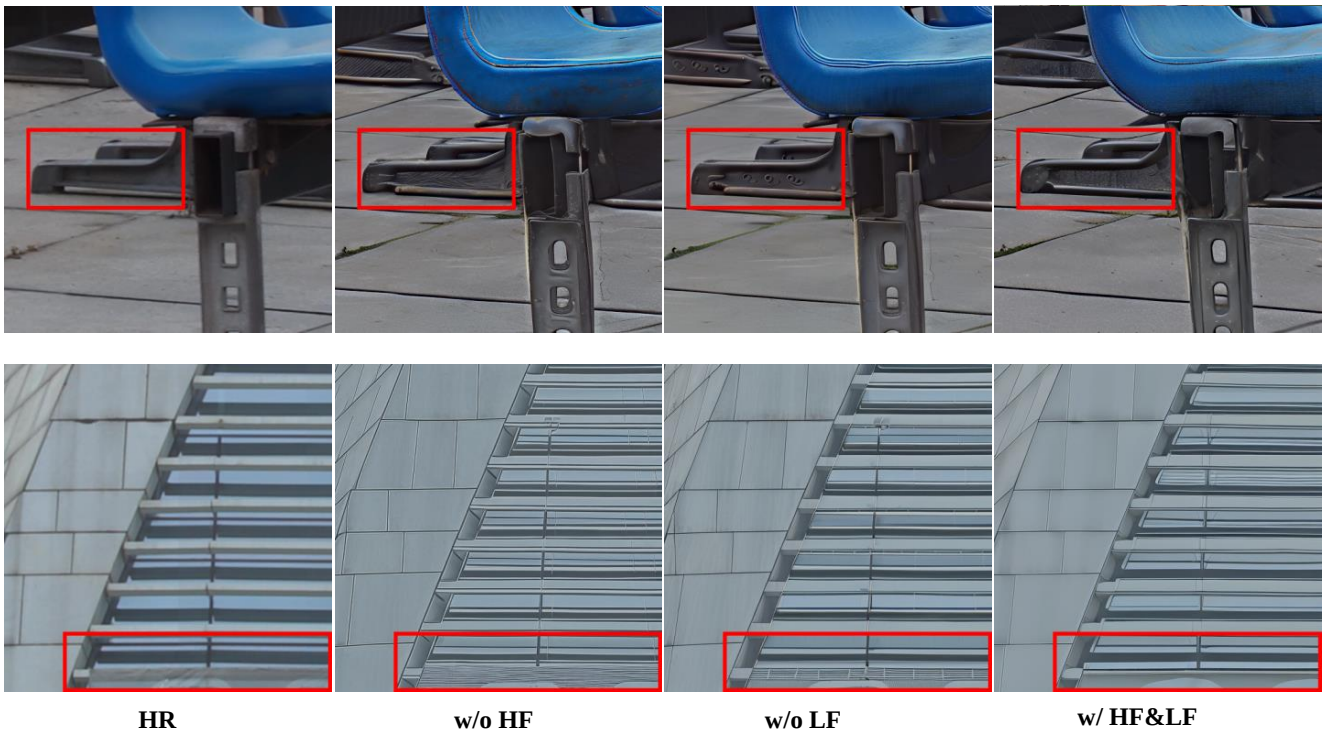


Figure 4. Visual comparison of different frequency-aware textual prior configurations. HF priors ensure correct fine textures, LF priors prevent hallucinations, and their combination yields the best results.

on both RealSR and DRealSR datasets in CLIP-IQA and MUSIQ, while maintaining competitive MANIQA scores. This indicates that initializing with global semantics provides a strong structural foundation, enabling local priors to operate more effectively in detail refinement without introducing artifacts.

In qualitative comparisons (Fig. 7), when the injection order changes from *Global* $\rightarrow$ *Local* to *Local* $\rightarrow$ *Global*, the super-resolved results exhibit noticeable artifacts and become blurry. This is because injecting local priors first forces the model to focus on object-level details before a coherent global structure is established, leading to misaligned textures and inconsistencies. In contrast, starting with global information provides a strong structural foundation, allowing subsequent local priors to refine details more accurately and consistently.

5.2.2. Effect of Low-High Frequency Prior Injection Order

We further conduct an ablation study to investigate the effect of the injection order of low-frequency (LF) and high-frequency (HF) textual priors in our framework. Table 7 reports the quantitative results on both the RealSR and DRealSR datasets. Across all perceptual metrics, our LF$\rightarrow$HF strategy consistently outperforms the HF$\rightarrow$LF variant, with notable gains in MANIQA and CLIP-IQA, indicating that establishing coarse structural semantics before refining textures leads to more accurate and perceptually pleasing reconstructions.

When the injection order is changed from LF$\rightarrow$HF to HF$\rightarrow$LF, the SR results exhibit noticeable degradation. As shown in the first row of Fig. 8, HF$\rightarrow$LF introduces extra, unrealistic textures, leading to less faithful reconstruction. In the second row, HF$\rightarrow$LF produces many yellow stain-like artifacts on the reconstructed surface. This degradation occurs because injecting high-frequency priors before establishing the correct low-frequency structure may guide the model to hallucinate details that are misaligned with the true scene. In contrast, our LF$\rightarrow$HF order first constrains the structural layout using low-frequency priors and then refines textures with high-frequency priors, resulting in more realistic and artifact-free outputs.

5.2.3. Effect of Guidance Scale

To evaluate the impact of guidance strength in our multi-branch Classifier-Free Guidance (CFG) strategy, we conduct an ablation study on the guidance scale parameter λ_s, which controls the amplification of conditional versus unconditional predictions during sampling. This scalar value affects the influence of textual priors on the denoising trajectory, and thus plays a critical role in balancing semantic alignment and generation stability.

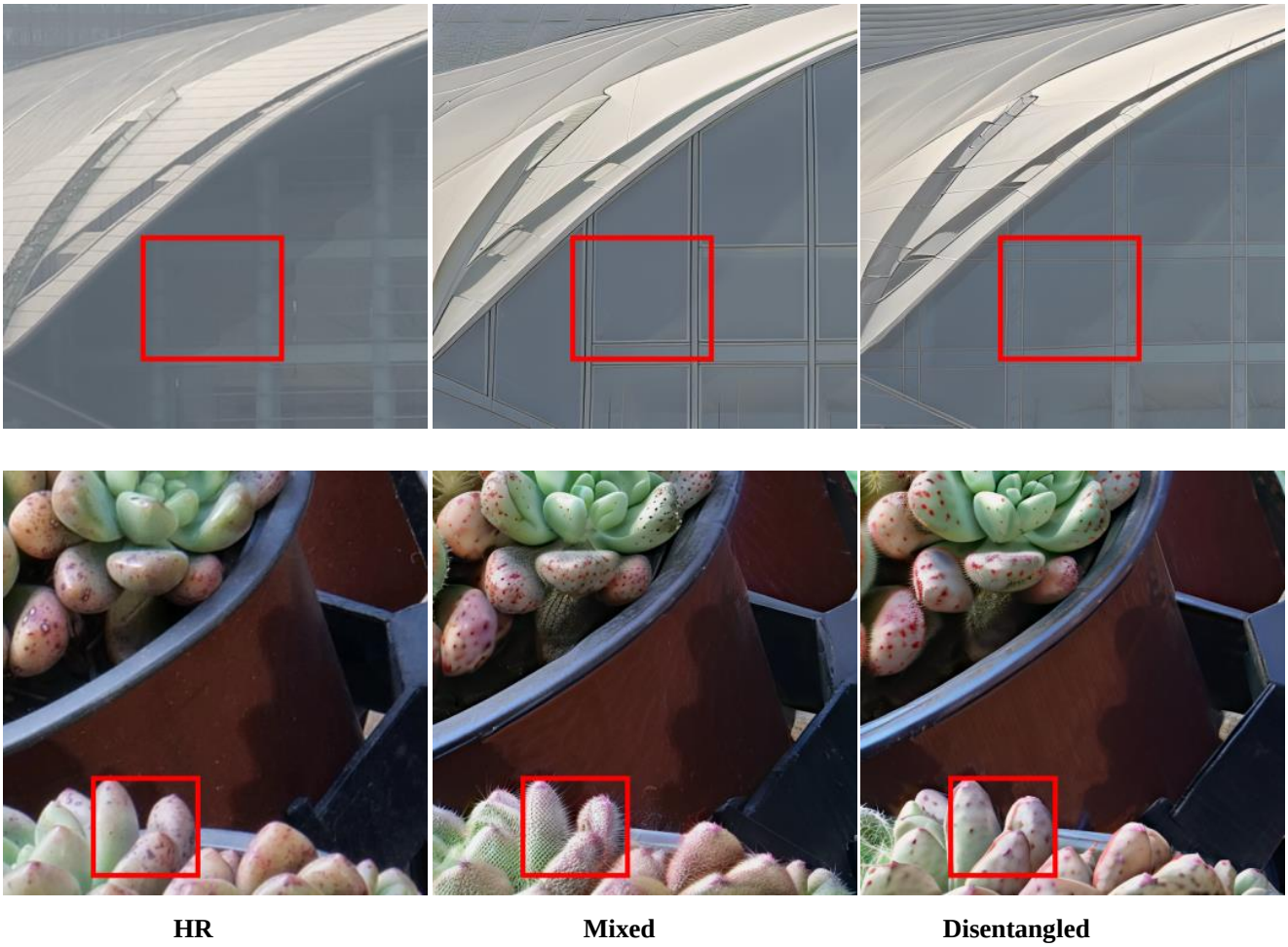


Figure 5. Visual comparison between mixed and disentangled frequency-aware learning. Disentangled modeling produces correct and realistic textures while avoiding hallucinated details.

We experiment with $\lambda_s \in [2.0, 10.5]$, increasing in steps of 0.5, and evaluate perceptual quality on the RealSR and DRealSR validation sets using three no-reference metrics: MUSIQ, MANIQA, and CLIP-IQA. As shown in Fig. 9, the performance trends exhibit consistent but nuanced behaviors across datasets and metrics:

- **MUSIQ**: Both datasets show steady gains as λ_s increases, with performance saturating between 7.0 and 9.0. This suggests that stronger guidance improves high-level aesthetic quality up to a point.
- **MANIQA**: The scores peak at $\lambda_s \approx 6.5$–7.0 and then decline, indicating that excessive guidance may impair structural consistency or introduce subtle artifacts.
- **CLIP-IQA**: This metric, which measures semantic-textual consistency, also peaks around $\lambda_s = 6.5$–7.5 before declining, likely due to over-conditioning causing oversharpening or hallucinated textures.

Importantly, we observe that overly small guidance values result in under-conditioning, leading to low-contrast, blurry outputs with weak semantic alignment. Conversely, excessively large values yield over-smoothed structures or unnatural textures due to exaggerated prompt interpretation.

Based on these observations, we select $\lambda_s = 7.0$ as the default setting for all experiments. It consistently delivers strong results across all three perceptual metrics and both datasets, offering a balanced trade-off between effective guidance and robust generation. This choice reflects the sweet spot where semantic priors meaningfully shape the output without destabilizing the denoising process.

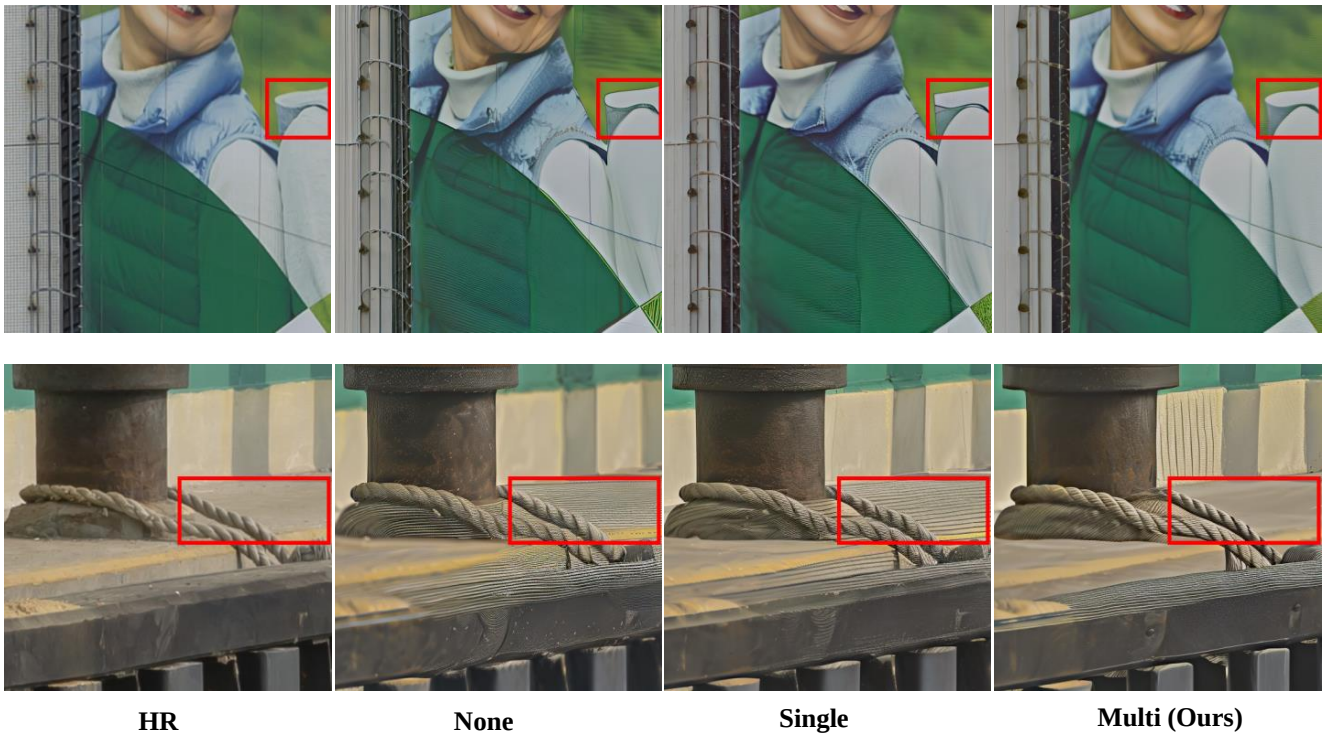


Figure 6. Visual comparison of different CFG strategies. Our multi-branch CFG produces more coherent global structure and finer details compared to no CFG and single-branch CFG.

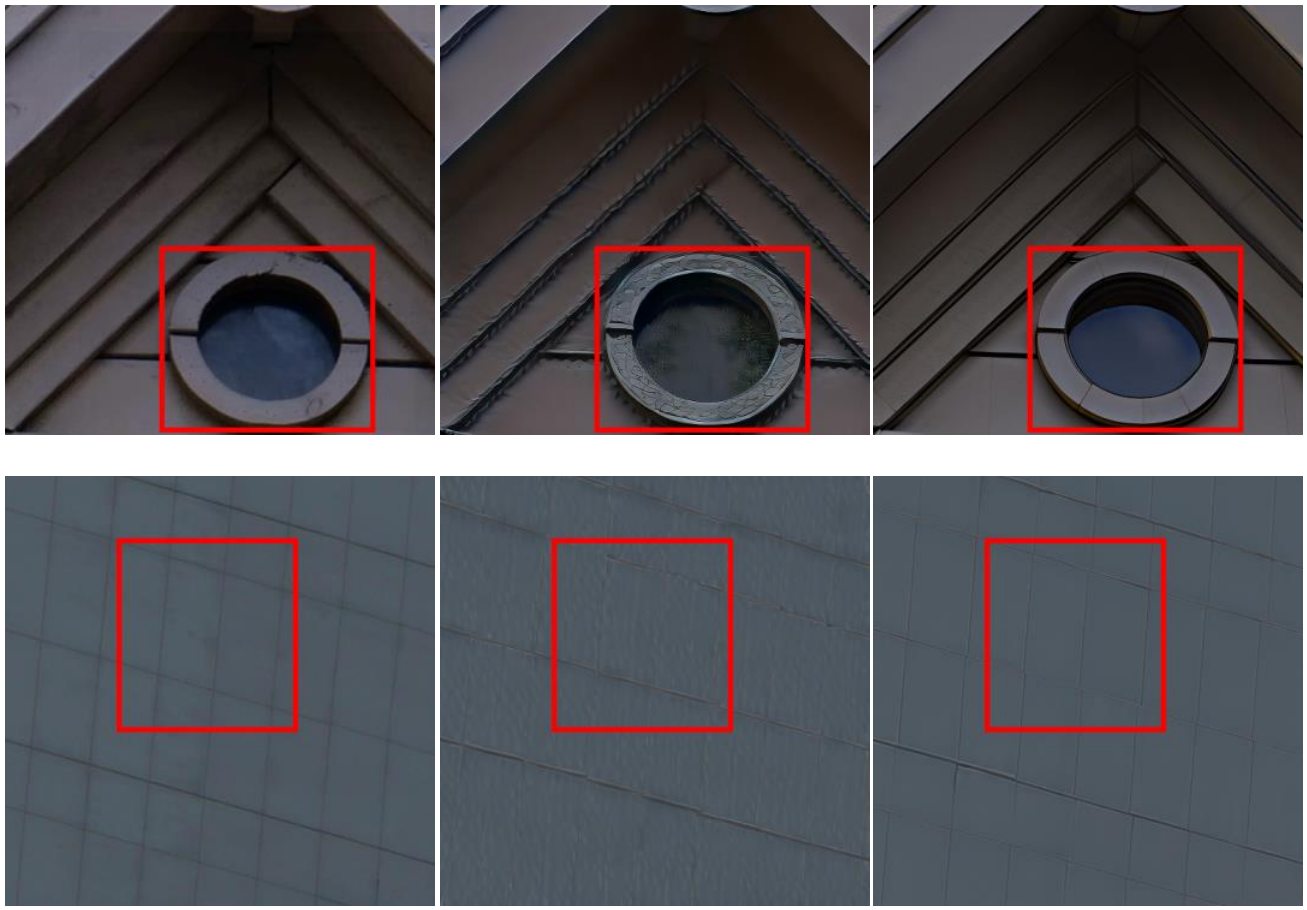
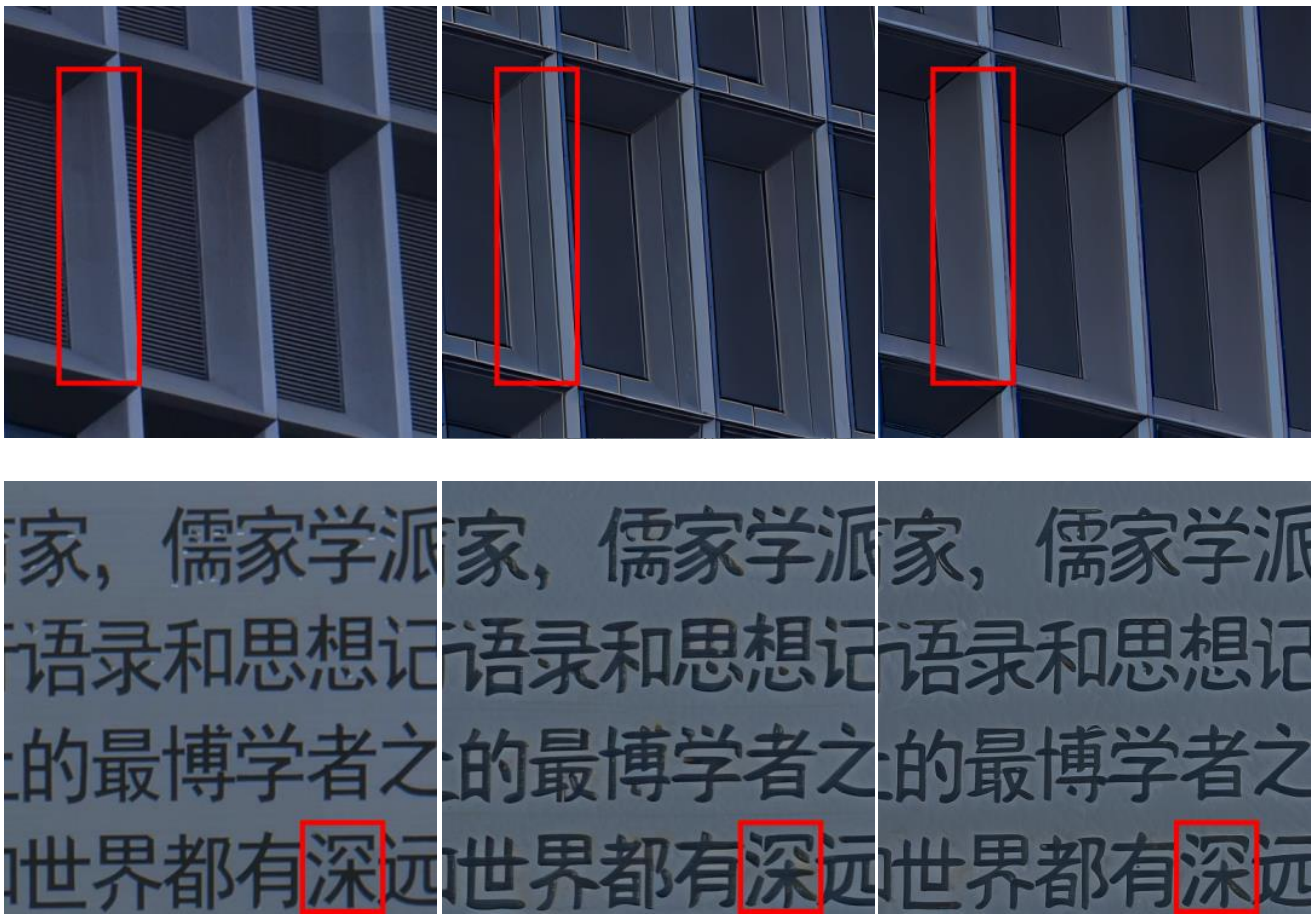


Figure 7. Qualitative comparison of different global–local prior injection orders. Changing the order from *Global* → *Local* to *Local* → *Global* results in more artifacts and blurriness, as local priors are injected before a coherent global structure is formed. Our *Global* → *Local* strategy first establishes the structural foundation, enabling more accurate and consistent detail refinement.



HR **HF→LF** **LF→HF（ours）**

Figure 8. Qualitative comparison of different LF–HF prior injection orders. LF→HF better preserves realistic textures without introducing artifacts.

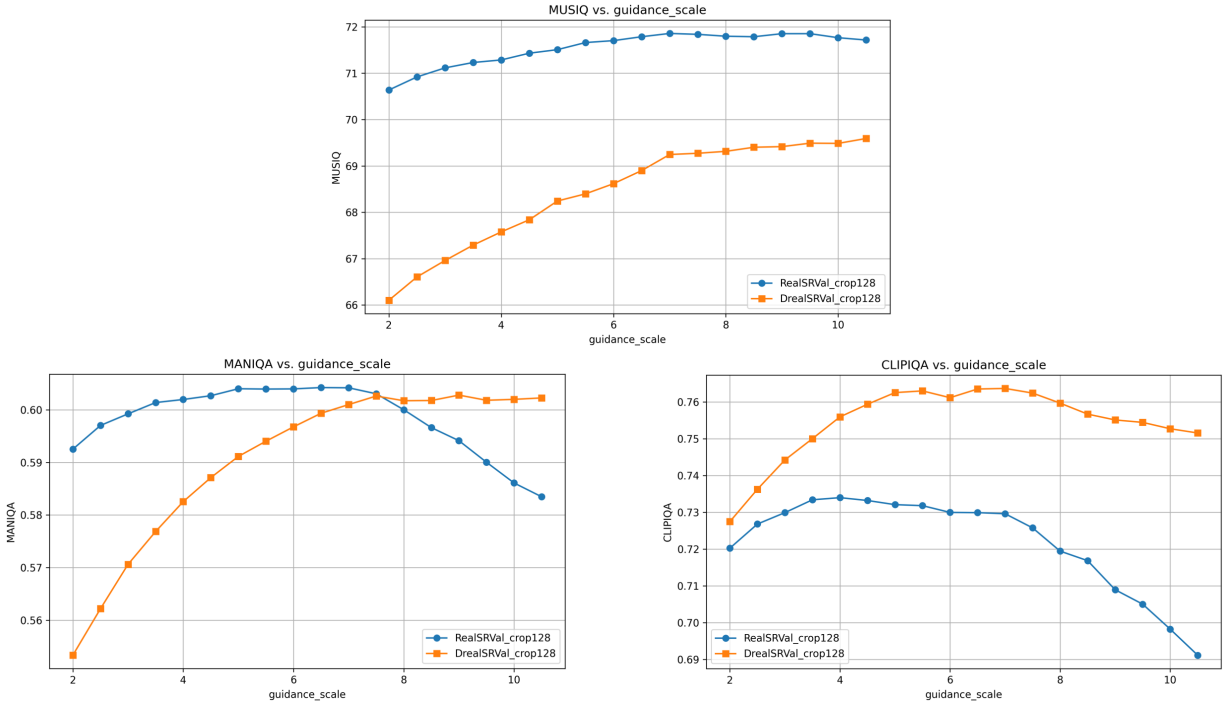


Figure 9. Effect of guidance scale on perceptual quality. We report MUSIQ, MANIQA, and CLIP-IQA scores on RealSR and DRealSR as λ_s increases.