TAPD-SR: BRIDGING THE DETAIL-CONDITIONING GAP FOR DIFFUSION-BASED FAITHFUL REAL-WORLD IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world image super-resolution (Real-ISR) using pretrained text-to-image (T2I) diffusion models achieves strong perceptual quality but often suffers from unfaithful textures. We trace this failure to a detail-conditioning gap: loss of low-resolution (LR) features in latent space, under-specified by coarse textual prompts, and insufficiently enforced along the denoising trajectory. To address this issue, we present TAPD-SR, a unified framework that strengthens LR-derived conditioning to improve generated texture fidelity. At the representation level, a Time-Aware Detail Feature Extractor (TDFE) captures high-frequency information directly from the LR input and adaptively injects it into the diffusion process according to the denoising timestep. At the semantic level, a Reasoning-driven Fine-grained Prompt Generator (RFPG) produces fine-grained global and object-level prompts from LR input using a fine-tuned multimodal large language model (MLLM). At the trajectory level, a Hierarchical Frequency Refinement Module (HFRM) extracts multi-scale frequency information from the LR input to predict a residual correction signal to the diffusion model's velocity output, providing additional structural and detail guidance for latent updates. Extensive experiments demonstrate that our TAPD-SR improves reconstruction fidelity while maintaining high perceptual quality.

1 INTRODUCTION

Real-ISR aims to reconstruct high-resolution (HR) images from low-resolution (LR) images that suffer from complex and random degradations. Recently, T2I diffusion models, such as Qwen-Image Wu et al. (2025) and Stable Diffusion Rombach et al. (2022), have attracted considerable attention due to their rich prior knowledge and remarkable ability to generate highly realistic textures and images. These models are now widely adopted in Real-ISR, substantially improving the perceptual quality of the reconstructed HR. However, these priors may also generate textures inconsistent with the LR input, resulting in artifacts and textures that do not exist in the LR, particularly for small text and intricate textures.

We attribute these problems to a detail-conditioning gap. The LR image must be $8\times$ downsampled and encoded into a latent space via a VAE encoder, which inherently discards high-frequency fine-grained features. When the T2I diffusion model encounters details that are weakened from the LR input, it relies on the textual prompt to guide the output. However, most existing Real-ISR methods only adopt simple prompts, for example, a global descriptive sentence and some object tagging labels. They provide limited detailed information at the input stage. As a result, the model is forced to infer textures based on its prior knowledge and LR's global semantics, producing outputs that are contextually coherent and visually plausible, yet often unfaithful to the actual LR content. Moreover, previous studies Yao et al. (2026); Park et al. (2026) have shown that visual and textual conditions injected into the diffusion backbone may be gradually weakened or lost in deeper MMDiT blocks. Consequently, the final velocity prediction at each denoising step may not sufficiently preserve LR-consistent structural and texture cues. When such prediction bias is repeatedly accumulated across sampling steps, the denoising trajectory may gradually deviate from the desired HR solution. Therefore, besides strengthening input and intermediate conditioning, it is necessary to directly refine the predicted velocity with reliable LR-derived cues.

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.

Based on the above analyses, we aim to improve reconstruction fidelity and maintain perceptual quality through detail-preserving visual conditioning, fine-grained prompt guidance, and denoising trajectory correction. In this paper, we propose TAPD-SR, a diffusion framework that bridges the detail-conditioning gap for faithful Real-ISR. To preserve more detailed features of LR, we employ a novel Time-Aware Detail Feature Extractor (TDFE) that utilizes lossless transformation and convolutions to extract high-frequency detail features and low-frequency features. The high-frequency features are then adaptively modulated via a time-aware modulation. Finally, the low-frequency features serve as structural cues to guide the image stream in retrieving and integrating the most relevant high-frequency details. We also incorporate a Reasoning-driven Fine-grained Prompt Generator (RFPG) to obtain fine-grained prompts. An MLLM first obtains degradation-aware reasoning ability by fine-tuning. Then we guide the fine-tuned MLLM to obtain detailed and dense global and object-level prompts on the LR image. The Two-level prompts enter the diffusion model through two independent branches because the native text stream has a limited effective context length for dense, fine-grained descriptions. In addition, to correct the denoising trajectory, we design a Hierarchical Frequency Refinement Module (HFRM). It further exploits structural and detail cues in the LR by decomposing the LR image into multiple scale-frequency components with nonlinear decomposition. Then, the effective information in each component and the current latent state are integrated to generate a correction signal to refine the velocity prediction.

Our main contributions are summarized as follows:

- We propose the TAPD-SR framework, which can generate HR with high fidelity and maintain high perceptual quality.
- We introduce three core components: TDFE, RPFG, and HFRM. They collaborate with each other at three levels to bridge the detail-conditioning gap and ultimately improve the fidelity of generated HR.
- Extensive experiments demonstrate that our method achieves good performance in both fidelity and perceptual quality. It has also reached SOTA level in text generation quality.

2 Related Works

To improve reconstruction fidelity, existing diffusion-based Real-ISR methods commonly inject LR information into the generation process. StableSR Wang et al. (2024) extracts timestep-aware LR latent features for feature modulation, while DiT4SR Duan et al. (2025) introduces an additional control stream into MMDiT for stronger LR conditioning. However, both rely primarily on VAE-compressed latent representations, where fine spatial details may already be weakened. VOSR Wu et al. (2026) introduces an external visual encoder to extract semantic embeddings from the LR image, providing semantic constraints, yet these semantic embeddings are too coarse and ambiguous for precise control. HDW-SR Yang et al. (2026) replaces the UNet-based diffusion framework with a wavelet-based diffusion network to preserve details, yet lacks text guidance and global modeling. In contrast, our TDFE explicitly extracts high-frequency information from the LR image before VAE encoding and adaptively injects it into the diffusion process.

To stimulate the generation ability of T2I diffusion models and control texture generation, many studies inject text prompts into diffusion models to enhance texture. SeeSR Wu et al. (2024) employs the advanced model RAM Zhang et al. (2024) to generate tags that describe the image content. FaithDiff Chen et al. (2025) relies on LLaVA Liu et al. (2023) to obtain a global description of the image. TSD-SR Dong et al. (2025) adopts a simpler strategy by using positive prompt terms, such as "shapely texture". These prompts mainly provide global semantics and offer limited guidance for specific local textures or textual content in the LR image. ODTSR Fang et al. (2026) improves text rendering by explicitly specifying textual content and its spatial location, yet its prompts are extracted from ground-truth (GT) images, and there are no detailed prompts for the text. Different from these methods, our RFPG automatically derives fine-grained prompts from the LR input, providing more precise textual conditioning for detailed textures and character reconstruction.

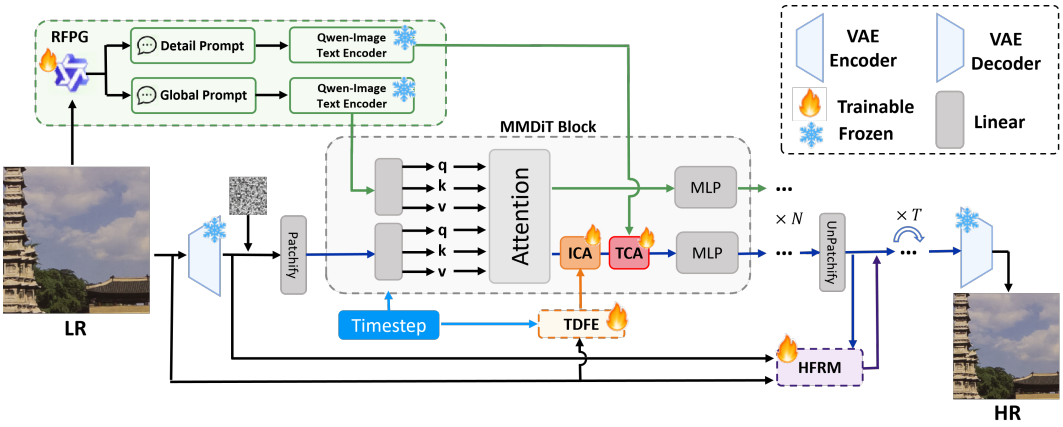


Figure 1: Overall pipeline of TAPD-SR. N and T denote the MMDiT block number and denoising timestep, respectively.

3 METHODOLOGY

As shown in Figure 1, TAPD-SR adopts Qwen-Image Wu et al. (2025) as the diffusion backbone. TDFE adaptively extracts detailed features and injects the diffusion process directly on LR input, whereas HFRM operates after MMDiT to directly correct the predicted velocity, thereby compensating for information that may be attenuated within deep MMDiT blocks. RFPG generates detailed prompts to fill in the gaps in detailed conditions at the semantic level. The following sections describe our core components in detail.

3.1 TIME-AWARE DETAIL FEATURE EXTRACTOR

To enhance the detailed features, we introduce a Time-Aware Detail Feature Extractor (TDFE) that explicitly preserves and injects high-frequency information in the LR image into the MMDiT image stream.

Given the LR image $\bar{\mathbf{x}}_L$, we first employ a Haar wavelet transform $\mathcal{H}$ to reduce the spatial resolution without discarding information, followed by $4\times$ pixel unshuffle PU_4 Shi et al. (2016). We can obtain four subbands, which can be expressed as:

$$[\mathbf{F}^{LL}, \mathbf{F}^{LH}, \mathbf{F}^{HL}, \mathbf{F}^{HH}] = \mathrm{PU}_4(\mathcal{H}(\bar{\mathbf{x}}_L)), \tag{1}$$

where the four subbands share the $8\times$ downsampling spatial resolution of the original input. $\mathbf{F}^{LL}$ mainly retains low-frequency structural information, whereas $\mathbf{F}^{LH}$, $\mathbf{F}^{HL}$, and $\mathbf{F}^{HH}$ contain complementary high-frequency detail information.

Since the prior knowledge and required detailed features activated at different timesteps t are different, we adaptively modulate the three high-frequency subbands according to the timestep embedding $\mathbf{e}_t$. Through a learnable projection layer $\mathcal{P}_b$, then we calculate the time-aware modulation parameters:

$$[\boldsymbol{\gamma}_t^b, \boldsymbol{\beta}_t^b] = \mathcal{P}_b(\mathbf{e}_t), b \in \{LH, HL, HH\} \tag{2}$$

The time-aware high-frequency features are obtained as:

$$\widetilde{\mathbf{F}}_t^b = (1 + \boldsymbol{\gamma}_t^b) \odot \mathrm{Norm}(\mathbf{F}^b) + \boldsymbol{\beta}_t^b. \tag{3}$$

The three modulated subbands are concatenated and processed by convolutions to form the detail representation $\mathbf{F}_t^H$, while $\mathbf{F}^{LL}$ is separately encoded by convolutions as a structural reference. Their spatial resolution after convolution matches the latent space.

We then tokenize the structural and detail representations using the same method as MMDiT, obtaining $\mathbf{C}^L$ and $\mathbf{C}_t^H$, respectively. We use an image cross-attention (ICA) to inject features. The MMDiT

image tokens are seen as queries, the low-frequency tokens as keys, and the high-frequency tokens as values:

$$\mathrm{ICA}(\mathbf{X}_{img}^{t}, \mathbf{C}^{L}, \mathbf{C}_{t}^{H}) = \mathrm{Softmax}\Big(\frac{(\mathbf{X}_{img}^{t} W_q)(W_k \mathbf{C}^{L})^{\top}}{\sqrt{d}}\Big) \times (W_v \mathbf{C}_{t}^{H}), \tag{4}$$

where W_q, W_k, W_v are learnable matrices. In this formula, low-frequency tokens as Keys provide query clues for the tokens in the MMDiT image stream to find the correct high-frequency information tokens.

3.2 Reasoning-driven fine-grained prompt generator

To extract fine-grained prompts, we divide the text prompt into a global prompt and an object-level prompt. We use Qwen3-VL Bai et al. (2025) as the MLLM. Appropriate instructions are employed to prompt Qwen3-VL Bai et al. (2025) to generate dense and detailed textual descriptions.

For the global prompt, we use the instruction: "Please detail describe the overall content and spatial organization of the image. Identify the scene, background, and objects, and describe their approximate locations, spatial relationships, and interactions. If it contains text, please recognize the text and describe the font, position, and color. For textual elements, if any characters are difficult to discern, please recover the correct text based on the recognizable characters and the contextual scene." Thus, we obtain a global textual description that can serve as a global prompt. This prompt is then encoded by the text encoder of Qwen-Image Wu et al. (2025) to yield the global text embedding $\mathbf{T}_{global}$, which is directly fed into the text stream of MMDiT.

For the object-level prompt, we use the instruction: "Please detail describe each object individually with emphasis on its local fine-grained visual appearance. For each object, specify its shape, color, material, surface texture, local patterns, edges, structural details, and other distinctive visual characteristics." Thus, we obtain detail textual description that can serve as an object-level prompt. This description is then encoded by the text encoder of Qwen-Image to yield the object-level text embedding $\mathbf{T}_{detail}$. $\mathbf{T}_{detail}$ is fused into the image stream of MMDiT via a text cross-attention (TCA). Specifically, the tokens in the MMDiT image stream are employed as queries, while the object-level text embedding serves as keys and values. The TCA is formulated as:

$$\mathrm{TCA}(\mathbf{X}_{\mathrm{img}}^{t}, \mathbf{T}_{\mathrm{detail}}) = \mathrm{Softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}\right) V, \tag{5}$$

where $Q = \mathbf{X}_{\mathrm{img}}^{t} W_Q, K = \mathbf{T}_{\mathrm{detail}} W_K, V = \mathbf{T}_{\mathrm{detail}} W_V$. W_Q, W_K, and W_V are learnable projection matrices.

To enable the MLLM to extract prompts from degraded LR inputs more accurately and comprehensively, we need to fine-tune the MLLM to acquire degradation-aware reasoning capabilities.

Specifically, given an HR image $\mathbf{x}_H$, we first use the frozen Qwen3-VL with the global and object-level instructions to generate the corresponding GT-derived descriptions, denoted as $\mathbf{y}_H^{\mathrm{G}}$ and $\mathbf{y}_H^{\mathrm{O}}$, respectively. These descriptions are regarded as GT prompts. During training, the LR image $\bar{\mathbf{x}}_L$ and the instructions are provided as inputs, while the GT-derived descriptions serve as supervision targets. The original Qwen3-VL parameters θ are frozen, and only trainable parameters $\Delta\theta$ are optimized by LoRA. We employ the standard autoregressive language modeling objective:

$$\mathcal{L}_{\mathrm{SFT}} = -\sum_{r \in \{\mathrm{G},\mathrm{O}\}} \frac{1}{\ell_r} \sum_{j=1}^{\ell_r} \log p_{\theta+\Delta\theta}\left(y_{H,j}^{r} \mid \bar{\mathbf{x}}_L, \mathbf{p}_r, \mathbf{y}_{H,<j}^{r}\right). \tag{6}$$

where $r \in \{\mathrm{G}, \mathrm{O}\}$ indexes the global and object-level descriptions, respectively. $\mathbf{p}_r$ is the instructions. $\mathbf{y}_H^r$ is the GT-derived pseudo-label, ℓ_r is its token length, $y_{H,j}^r$ is its j-th token, and $\mathbf{y}_{H,<j}^r$ denotes the preceding target tokens used for teacher forcing. $p_{\theta+\Delta\theta}$ denotes the conditional next-token probability of the fine-tuned model. The loss is computed only over target response tokens and averaged over training examples.

Through GT-derived supervision, the fine-tuned Qwen3-VL learns to generate degradation-aware and visual descriptions from degraded LR input, thereby improving the reliability of the generated prompts.

3.3 HIERARCHICAL FREQUENCY REFINEMENT MODULE

Complex degradations in the LR observation can obscure reliable structural and texture information. A previous study Cui et al. (2025); Chen et al. (2024) finds that different degradations affect image components across scales to varying degrees. Therefore, rather than treating the LR observation as a unified representation, we decompose it into multiple nonlinear scale components to separately exploit structural and detail information at different scales. Based on this motivation, we propose HFRM, which decomposes the LR input into multiple scale-frequency components before VAE encoding to refine the diffusion velocity prediction.

Given the LR image input $\bar{\mathbf{x}}_L$, we employ the spectral total variation (TV) decomposition Gilboa (2014) to separate it into five complementary scale-frequency components without changing resolution. Let $\Phi(\bar{\mathbf{x}}_L)$ denote its TV spectral representation and $\{\tau_i\}_{i=0}^{5}$ denote scale boundaries. The i-th component is obtained by:

$$\mathbf{F}_i = \int_{\tau_{i-1}}^{\tau_i} \Phi(\bar{\mathbf{x}}_L; s)\, ds, \qquad i = 1, \ldots, 5, \tag{7}$$

Here, s denotes the continuous TV-flow scale. We uniformly partition the TV scale interval $[0, S]$ into five subintervals, with boundaries $\tau_i = iS/5$. The finite-scale reconstruction residual is added to the last component. Other details and related experimental analysis are included in our appendix. We then encode each component and the complete LR input using the pretrained VAE encoder, which can be expressed by $\mathbf{z}_i^f = E_{\mathrm{VAE}}(\mathbf{F}_i), \mathbf{z}_L = E_{\mathrm{VAE}}(\bar{\mathbf{x}}_L)$.

Since an isolated frequency component may lose correct global structural and semantic context after VAE encoding, we use the complete LR latent $\mathbf{z}_L$ as a global anchor for each frequency branch $\mathbf{h}_i = \mathrm{SiLU}\left(\phi_i^{1\times1}\left([\mathbf{z}_i^f \| \mathbf{z}_L]\right)\right)$.

The five frequency features are aggregated into $\mathbf{z}_F$ and the initial velocity prediction $\mathbf{v}_G^t$ is produced by MMDiT at timestep t, which can be expressed as follows:

$$\mathbf{z}_F = \mathrm{SiLU}\left(\phi_F^{1\times1}([\mathbf{h}_1 \| \cdots \| \mathbf{h}_5])\right), \quad \mathbf{h}_R^t = \mathrm{SiLU}\left(\phi_{R,1}^{1\times1}([\mathbf{z}_F \| \mathbf{v}_G^t])\right), \tag{8}$$

where ϕ and $\|$ denote convolution operations and concatenation, respectively. Then, we can obtain the velocity correction signal via convolution operations. This processing can be expressed as follows:

$$\Delta\mathbf{v}_t = \phi_{R,2}^{1\times1}(\mathbf{h}_R^t), \quad \mathbf{v}_R^t = \mathbf{v}_G^t + \Delta\mathbf{v}_t, \tag{9}$$

where $\Delta\mathbf{v}_t$ is the velocity correction signal, and $\mathbf{v}_R^t$ is the refined velocity prediction. During training, the flow matching loss is computed on $\mathbf{v}_R^t$. During inference, HFRM refines the velocity at each sampling step before the sampler updates the latent state.

3.4 TRAINING OBJECTIVE

Following Qwen-Image Wu et al. (2025), we adopt a conditional flow matching objective. Given an GT image $\mathbf{x}_G$ and its paired LR image $\bar{\mathbf{x}}_L$, we obtain the GT latent using the frozen VAE encoder: $\mathbf{z}_G = E_{\mathrm{VAE}}(\mathbf{x}_G)$. We sample Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a timestep $t \sim p(t)$, where $p(t)$ denotes the training timestep distribution. The intermediate latent is constructed as:

$$\mathbf{z}_t = t\mathbf{z}_G + (1-t)\boldsymbol{\epsilon}. \tag{10}$$

The corresponding target velocity is $\mathbf{z}_G - \boldsymbol{\epsilon}$.

Conditioned on the LR observation and RFPG text embeddings, MMDiT produces an initial velocity prediction, which is further refined by HFRM. We supervise the refined prediction $\mathbf{v}_R^t$ using:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{(\mathbf{x}_G, \bar{\mathbf{x}}_L)\sim\mathcal{D},\, t\sim p(t),\, \boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left\|\mathbf{v}_R^t - (\mathbf{z}_G - \boldsymbol{\epsilon})\right\|_2^2\right], \tag{11}$$

where $\mathcal{D}$ denotes the paired training dataset. During training, the pretrained Qwen-Image backbone remains frozen, while TDFE, TCA, and HFRM are jointly optimized through this objective.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. We trained our Real-ISR method on a mixed dataset of Flickr2K Timofte et al. (2017), LSDIR Li et al. (2023), and the first 10K face images from FFHQ Karras et al. (2019). We use the same degradation pipeline as Real-ESRGAN Wang et al. (2021) to generate HR-LR image pairs in the above mixed dataset. The resolution is 128×128 for LR images and 512×512 for HR images. In this work, we mainly focus on ×4 Real-ISR tasks.

In order to fully test the performance of various Real-ISR methods, we use DIV-2K Agustsson & Timofte (2017), DrealSR Wei et al. (2020), and RealSR Cai et al. (2019) for testing. Moreover, UZ-ST Luo et al. (2026) is used to test the text generation performance. This dataset contains a substantial amount of both Chinese and English text with real-world degradations. All experiments are conducted with a scaling factor of ×4. DrealSR and RealSR, respectively, consist of 93 and 100 images. Following SeeSR Wu et al. (2024), changing image size is adopted for these two datasets, and the resolution of LR images is set to 128×128. DIV-2K is a synthetic dataset that utilizes Real-ESRGAN's degradation pipeline to artificially synthesize LR images.

Implementary detail. We divide the training into two stages. In the first stage, we fine-tune the MLLM using LoRA with a rank of 64 on 8 NVIDIA A6000 GPUs. The specific MLLM version is Qwen3-VL-8B-Instruct, with a batch size of 64 and a learning rate of 5e-5. The optimization objective is $\mathcal{L}_{\mathrm{SFT}}$.

In the second stage, we train our TAPD-SR. We freeze the pretrained Qwen-Image Wu et al. (2025) backbone and only optimize the trainable parameters. The training runs for 20K iterations with a batch size of 32 and a learning rate of 5e-5, using the AdamW Loshchilov & Hutter (2017) optimizer. The number of denoising diffusion steps is set to 20. The optimization objective is $\mathcal{L}_{\mathrm{FM}}$.

Evaluation metrics. To ensure a thorough and comprehensive evaluation of various methods' performance, we utilize a series of reference and no-reference evaluation metrics. Among these, PSNR and SSIM are reference-based metrics for assessing fidelity. On the other hand, LPIPS Zhang et al. (2018) and DISTS Ding et al. (2020) serve as reference-based metrics of perceptual quality. MANIQA Yang et al. (2022), MUSIQ Ke et al. (2021), and CLIPIQA Wang et al. (2023) are adopted as no-reference metrics for image quality assessment.

Additionally, we evaluate text recognition quality using OCR-A, which measures exact-match recognition accuracy, and NED, which measures edit-based text similarity. Their definitions are provided in the appendix.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Quantitative comparison results are shown in Table 1. TAPD-SR achieves a favorable fidelity-perception balance across the three datasets. It consistently obtains the best LPIPS and DISTS, while achieving the best MUSIQ on the three datasets. In addition, our method ranks first in SSIM on DIV-2K and RealSR and maintains competitive PSNR, demonstrating strong reconstruction fidelity. We also achieve competitive no-reference perceptual scores on MUSIQ and CLIPIQA on the three datasets. These results indicate that our method achieves competitive performance on both fidelity and perceptual quality.

We further conduct separate evaluations of text-generation capability on the UZ-ST dataset. The results are shown in Table 2. Our method obtains top-ranked performance on both OCR-A and NED, yielding the overall best results. This validates the superiority of our approach over existing SOTA methods in text generation. This good performance mainly comes from our RFPG and Qwen-Image, as Qwen-Image has stronger support for Chinese than Stable Diffusion Rombach et al. (2022) and other T2I diffusion models.

For qualitative comparison, Figure 2 presents qualitative comparisons on challenging real-world scenes containing text and complex textures. In the signboard example, several competing methods fail to preserve the correct character structures and introduce noticeable deformation or texture artifacts. In contrast, our method better achieves the restoration of text and typography with fewer artifacts or structural distortions. In the Tokyo Tower example, we can see that competing meth-

Table 1: Quantitative comparison on DIV-2K, DrealSR, and RealSR. The best and second-best results are highlighted in **bold** and underline, respectively.

Dataset	Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MANIQA↑	MUSIQ↑	CLIPIQA↑
DIV-2K	ODTSR	24.0846	0.6138	0.2830	0.1896	0.6270	68.5467	0.6718
	DiT4SR	22.6237	0.5683	0.4194	0.2710	0.5285	63.7401	0.5889
	FaithDiff	23.3027	0.5690	0.3180	0.2071	**0.6802**	69.5080	0.6681
	SeeSR	23.6810	0.6049	0.3193	0.1965	0.6232	68.6969	0.6948
	TSD-SR	23.0215	0.5808	0.2673	0.1821	0.6226	71.6930	**0.7415**
	VOSR	**24.1469**	0.6055	0.2618	0.1869	0.5976	67.4545	0.5757
	Ours	24.1235	**0.6363**	**0.2486**	**0.1712**	0.6221	**71.7321**	0.7128
DrealSR	ODTSR	**28.1977**	0.7706	0.2912	0.2108	0.6285	67.0321	0.7135
	DiT4SR	27.0880	0.7317	0.3834	0.2700	0.5121	60.1390	0.5567
	FaithDiff	27.5319	0.7228	0.3625	0.2485	**0.6375**	64.7283	0.6259
	SeeSR	28.0728	0.7684	0.3175	0.2315	0.6044	65.0775	0.6908
	TSD-SR	27.7671	0.7559	0.2967	0.2136	0.5865	66.6064	**0.7344**
	VOSR	28.1711	**0.7782**	0.2841	0.2062	0.4738	52.4059	0.4545
	Ours	28.1867	0.7755	**0.2689**	**0.1829**	0.6198	**67.5667**	0.7211
RealSR	ODTSR	25.3641	0.7279	0.2743	0.2003	**0.6717**	70.5623	0.6880
	DiT4SR	24.0509	0.6926	0.3451	0.2525	0.5777	65.2784	0.5689
	FaithDiff	**25.4159**	0.7159	0.2869	0.2122	0.6699	68.5656	0.6137
	SeeSR	25.1486	0.7211	0.3007	0.2224	0.6439	69.8174	0.6699
	TSD-SR	24.8065	0.7172	0.2731	0.2105	0.6329	71.1811	0.7162
	VOSR	25.4102	0.7597	0.2768	0.2051	0.5638	60.1911	0.4569
	Ours	25.3913	**0.7721**	**0.2516**	**0.1817**	0.6601	**71.4382**	**0.7258**

Table 2: Quantitative comparison of text accuracy on the UZ-ST dataset. The best results are highlighted in **bold**.

Method	ODTSR	DiT4SR	FaithDiff	SeeSR	TSD-SR	VOSR	Ours
OCR-A ↑ (%)	37.2	23.7	27.9	22.8	26.6	32.6	**48.8**
NED ↑	0.6174	0.5391	0.5522	0.5259	0.5428	0.5892	**0.6921**

ods either blur the fine lattice structures or generate irregular high-frequency patterns. TAPD-SR better preserves the geometric structure and texture details. These results demonstrate that the proposed components effectively improve reconstruction fidelity while retaining perceptually realistic textures.

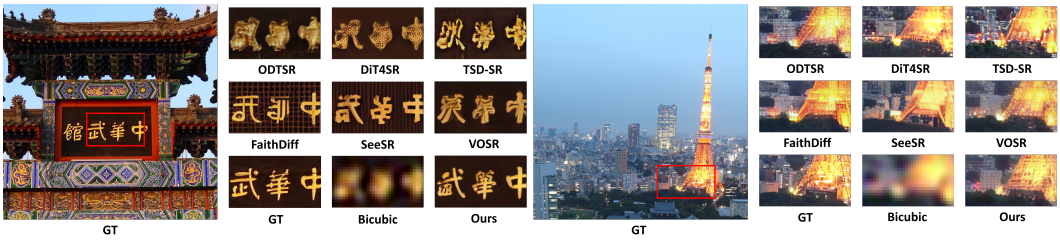


Figure 2: The visual results of qualitative comparison with other SOTA Real-ISR methods.

5 Ablation Study

5.1 The effectiveness of TDFE

We investigate the effectiveness of different designs in TDFE. As shown in Table 3, removing TDFE noticeably degrades all metrics, especially LPIPS and DISTS, demonstrating the importance of explicitly preserving high-frequency information from the LR input. Removing the time-aware modulation (TAM) also leads to consistent performance degradation, indicating that the demand for fine

details varies across denoising timesteps and should be adaptively controlled. Replacing ICA with simple zero-convolution further degrades the results, showing that selective feature interaction is more effective than direct feature aggregation. Moreover, using high-frequency information tokens themselves as keys performs worse than using low-frequency information tokens, validating that structural information provides a reliable cue for locating and injecting appropriate high-frequency details. The complete TDFE achieves the best overall performance, confirming the effectiveness of our design.

From Figure 3, it can be seen that ours exhibits the closest texture and detail to the GT image, while still maintaining excellent perceptual quality. The other variants all suffer from texture loss and hallucinated texture generation. This clearly demonstrates the effectiveness of our TDFE.

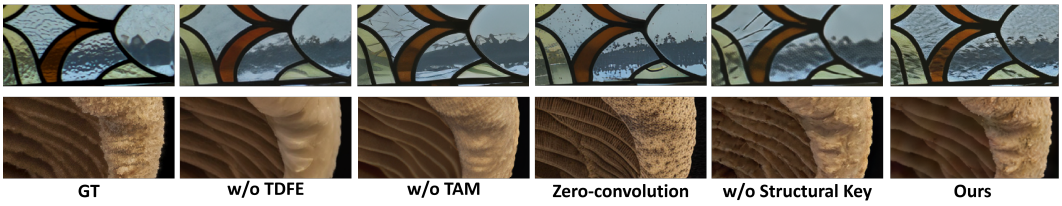


Figure 3: Visual results of various variants in the TDFE ablation study.

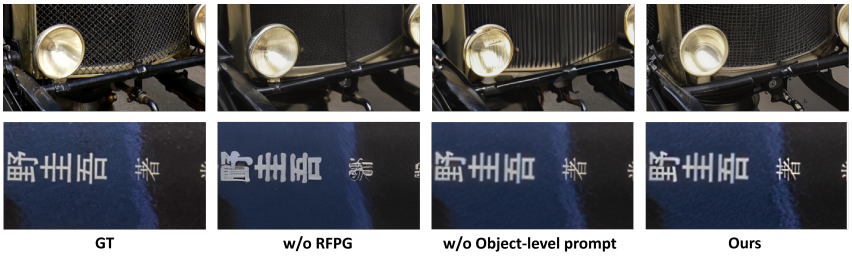


Figure 4: Visual results of various variants in the RFPG ablation study.

Table 3: Ablation study of the Time-Aware Detail Feature Extractor (TDFE) on RealSR. TAM denotes the time-aware modulation. The best results are highlighted in **bold**.

Variant	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑	CLIPIQA↑
w/o TDFE	0.7437	0.2689	0.1996	70.4641	0.6981
w/o TAM	0.7578	0.2601	0.1916	70.8234	0.7105
Zero-convolution	0.7474	0.2649	0.1951	70.7534	0.7067
w/o Structural Key	0.7631	0.2574	0.1869	71.0233	0.7012
Full (Ours)	**0.7721**	**0.2516**	**0.1817**	**71.4382**	**0.7258**

5.2 The effectiveness of RFPG

We evaluate the effectiveness of RFPG by progressively introducing global and object-level prompt guidance. As shown in Table 4, removing RFPG results in clear degradation in perceptual quality, indicating that RFPG is effective. Introducing the global prompt improves no-reference metrics by providing richer global semantics. Adding the object-level prompt without TCA further improves both no-reference perceptual metrics and reference-based perceptual metrics, demonstrating the benefit of explicitly describing object-level textures and characteristics. Finally, separately injecting the object-level prompt into the image stream through TCA achieves the best performance. This verifies that the object-level prompt can provide more effective guidance when explicitly interacting with image tokens, rather than being treated only as global textual conditioning. This is also because text streams cannot accept tokens that are too long, and our dense prompts exceed the maximum supported token length. Therefore, injecting them separately works better.

In Figure 4, we can see that the global prompt is helpful for text recovery and texture generation. But without finer object-level prompts, unfaithful textures may still be generated. After adding fine object-level prompts, the generated texture is closest to the GT image. This intuitively demonstrates the effectiveness of each of our designs in RFPG.

Table 4: Ablation study of the Reasoning-driven Fine-grained Prompt Generator (RFPG) on RealSR. The best results are highlighted in **bold**.

Variant	MANIQA↑	MUSIQ↑	CLIPIQA↑	LPIPS↓	DISTS↓
w/o RFPG	0.6441	70.2864	0.6918	0.2621	0.1938
Global Prompt	0.6518	70.8653	0.7029	0.2585	0.1907
w/o TCA	0.6557	71.0121	0.7147	0.2583	0.1878
Full RFPG	**0.6601**	**71.4382**	**0.7258**	**0.2516**	**0.1817**

5.3 The effectiveness of HFRM

We evaluate the effectiveness of HFRM and investigate its sensitivity to the number of scale-frequency components on RealSR. As shown in Table 5, the default five-component configuration improves PSNR and SSIM, compared with removing HFRM. It also reduces LPIPS and DISTS, while increasing MUSIQ. These results support the effectiveness of HFRM in improving reconstruction quality. We vary the number of scale-frequency components from one to seven. As shown in Figure 5, using a single component underperforms on most metrics, while increasing the number of components from one to five shows an overall improving trend across the evaluated metrics. Further increasing it to six or seven slightly improves SSIM but degrades the other metrics relative to five components. We therefore adopt five components as the default, as it achieves the best PSNR, LPIPS, DISTS, and MUSIQ among the evaluated configurations.

Table 5: The results of the ablation study of HFRM on RealSR. The best results are highlighted in **bold**.

Metric	w/o HFRM	1 level	2 levels	3 levels	4 levels	5 levels	6 levels	7 levels
PSNR↑	25.1472	24.9841	25.2145	25.3028	25.3640	**25.3913**	25.3903	25.3876
SSIM↑	0.7395	0.7336	0.7463	0.7538	0.7587	0.7721	0.7738	**0.7746**
LPIPS↓	0.2696	0.2723	0.2643	0.2587	0.2553	**0.2516**	0.2541	0.2562
DISTS↓	0.1915	0.1905	0.1964	0.1908	0.1885	**0.1817**	0.1870	0.1887
MUSIQ↑	70.9481	70.9010	70.8542	71.0026	71.3180	**71.4382**	71.4105	71.4043

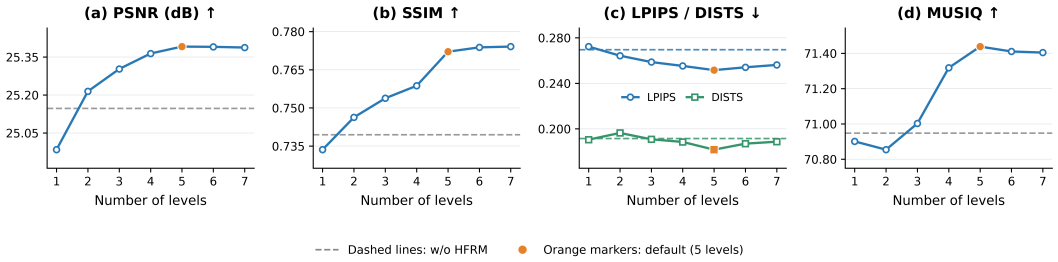


Figure 5: Effect of the number of scale-frequency components in HFRM on RealSR.

6 Conclusion

We propose TAPD-SR, a diffusion-based framework for Real-ISR. By combining timestep-aware detail injection, fine-grained prompts, and hierarchical frequency refinement, TAPD-SR bridges the gap in detail conditions. Extensive experiments demonstrate that it achieves excellent fidelity and perceptual quality with improved text generation performance. **More analysis and details can be found in our appendix.**

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing research methodology or experiments, implementing methods, cleaning and reformatting datasets, or interpreting results; qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for editing the research paper to improve readability, format references, and create or edit software code. We have reviewed all AI-assisted work: all suggested text modifications and code snippets were manually evaluated, tested, and verified for correctness and originality by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1122–1131. IEEE, 2017.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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. IEEE, 2019.

Genggeng Chen, Kexin Dai, Kangzhen Yang, Tao Hu, Xiangyu Chen, Yongqing Yang, Wei Dong, Peng Wu, Yanning Zhang, and Qingsen Yan. Bracketing image restoration and enhancement with high-low frequency decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6097–6107. IEEE, 2024.

Junyang Chen, Jinshan Pan, and Jiangxin Dong. Faithdiff: Unleashing diffusion priors for faithful image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28188–28197. IEEE, 2025.

Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Khan. Adair: Adaptive all-in-one image restoration via frequency mining and modulation. In *International Conference on Learning Representations*, volume 2025, pp. 101306–101327, 2025.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020.

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. IEEE, 2025.

Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S Ren, Chunle Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18948–18958, 2025.

Yushun Fang, Yuxiang Chen, Shibo Yin, Qiang Hu, Jiangchao Yao, Ya Zhang, Xiaoyun Zhang, and Yanfeng Wang. One-step diffusion transformer for controllable real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23440–23450, 2026.

Guy Gilboa. A total variation spectral framework for scale and texture analysis. *SIAM Journal on Imaging Sciences*, 7(4):1937–1961, 2014.

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. 5128–5137. IEEE, 2021.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1775–1787, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Minxing Luo, Linlong Fan, Qiushi Wang, Ge Wu, Yiyan Luo, Yuhang Yu, Jinwei Chen, Yaxing Wang, Qingnan Fan, and Jian Yang. Restore text first, enhance image later: Two-stage scene text image super-resolution with glyph structure guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30553–30563, 2026.

Dogyun Park, Moayed Haji-Ali, Yanyu Li, Willi Menapace, Sergey Tulyakov, Hyunwoo J Kim, Aliaksandr Siarohin, and Anil Kag. Sprint: Sparse-dense residual fusion for efficient diffusion transformers. 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10674–10685. ieee, 2022.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, 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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883, 2016.

Jean-Luc Starck, Jalal Fadili, and Fionn Murtagh. The undecimated wavelet decomposition and its reconstruction. *IEEE Transactions on Image Processing*, 16(2):297–309, 2007.

Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 114–125, 2017.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023.

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*, 2024.

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*, pp. 1905–1914, 2021.

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 *European Conference on Computer Vision*, pp. 101–117. Springer, 2020.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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, 2024.

Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, Xiangtao Kong, Jixin Zhao, Shihao Wang, and Lei Zhang. Vosr: A vision-only generative model for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16311–16321, June 2026.

Chao Yang, Boqian Zhang, Jinghao Xu, and Guang Jiang. Hdw-sr: High-frequency guided diffusion model based on wavelet decomposition for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23462–23472, 2026.

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*, pp. 1191–1200, 2022.

Yuxuan Yao, Yuxuan Chen, Hui Li, Kaihui Cheng, Qipeng Guo, Yuwei Sun, Zilong Dong, Jingdong Wang, and Siyu Zhu. Prompt reinjection: Alleviating prompt forgetting in multimodal diffusion transformers. In *International Conference on Machine Learning (ICML)*, 2026.

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

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1724–1732. IEEE, 2024.

A Appendix

A.1 Text Recognition Metrics

We evaluate text recognition quality using two complementary metrics: exact-match recognition accuracy (OCR-A) and edit-based text similarity (NED).

OCR-A. OCR-A measures the percentage of evaluated images for which the recognized text exactly matches the ground-truth transcription after normalization. For the i-th image, let p_i denote the text sequence obtained by concatenating the recognized text segments returned by the OCR engine, and let a_i denote the corresponding ground-truth annotation. The normalization function $\mathcal{N}(\cdot)$ removes whitespace and converts text to lowercase. OCR-A is computed as

$$\text{OCR-A} = \frac{100}{N} \sum_{i=1}^{N} \mathbf{1}\left[\mathcal{N}(p_i) = \mathcal{N}(a_i)\right], \tag{12}$$

where N is the number of evaluated images with available ground-truth text annotations, and $\mathbf{1}[\cdot]$ is the indicator function. OCR-A is reported as a percentage, with higher values indicating better recognition accuracy. This metric assigns no partial credit to an incorrect transcription.

NED. Let p denote a text sequence recognized from a super-resolved image and g its corresponding reference sequence. We use $|p|$ and $|g|$ to denote their lengths, and $d(p, g)$ to denote the Levenshtein edit distance with unit insertion, deletion, and substitution costs. For a text pair with $\max(|p|, |g|) > 0$, NED is computed as

$$\text{NED}(p, g) = 1 - \frac{d(p, g)}{\max(|p|, |g|)}. \tag{13}$$

Following the terminology used in prior work, we refer to this metric as NED. Since the normalized edit distance is subtracted from one, the resulting score measures text similarity, with higher values indicating closer agreement between the recognized and reference text sequences. The reference sequence g is obtained by applying OCR to the corresponding ground-truth image.

A.2 Details of Spectral Total Variation Decomposition

We elaborate on the spectral total variation (TV) decomposition used in HFRM, following Gilboa Gilboa (2014). We use s for the TV-flow scale to distinguish it from the diffusion timestep t.

TV flow. For clarity, consider a scalar image $f : \Omega \to \mathbb{R}$ on the spatial domain Ω. The isotropic TV functional is formally written as

$$J_{\text{TV}}(u) = \int_{\Omega} \|\nabla u(\boldsymbol{\xi})\|_2 \, d\boldsymbol{\xi}, \tag{14}$$

where $\boldsymbol{\xi}$ denotes spatial position. The corresponding TV flow is defined by

$$\frac{\partial u(s)}{\partial s} = -p(s), \qquad p(s) \in \partial J_{\text{TV}}(u(s)), \qquad u(0) = f, \tag{15}$$

where ∂J_{TV} denotes the subdifferential of the TV functional. Under homogeneous Neumann boundary conditions, the flow preserves the image mean.

Spectral representation. The TV spectral representation is defined as

$$\phi_f(s) = s \frac{\partial^2 u(s)}{\partial s^2}, \tag{16}$$

with derivatives interpreted weakly when necessary. Unlike Fourier frequency, the TV scale depends on both spatial extent and contrast. The resulting components therefore characterize nonlinear image scales rather than fixed Fourier frequency bands.

Finite-scale reconstruction. For a finite stopping scale $S > 0$, define

$$r_S = u(S) - S\frac{\partial u(S)}{\partial s}. \tag{17}$$

Integration by parts gives

$$\begin{aligned}\int_0^S \phi_f(s)\,ds &= \int_0^S s\,\frac{\partial^2 u(s)}{\partial s^2}\,ds \\ &= S\frac{\partial u(S)}{\partial s} - u(S) + f.\end{aligned} \tag{18}$$

Here, the boundary term at zero is understood under the corresponding regularity assumptions. Consequently,

$$f = \int_0^S \phi_f(s)\,ds + r_S. \tag{19}$$

The residual retains the contribution not included in the truncated spectral integral.

Component construction in HFRM. Let $\Phi(\bar{\mathbf{x}}_L; s)$ denote the image-valued spectral representation of the LR input, and let $\mathbf{r}_S$ be its reconstruction residual. We partition $[0, S]$ into K intervals with

$$\tau_i = \frac{iS}{K}, \qquad i = 0, \ldots, K. \tag{20}$$

The components are defined by

$$\mathbf{F}_i = \begin{cases} \displaystyle\int_{\tau_{i-1}}^{\tau_i} \Phi(\bar{\mathbf{x}}_L; s)\,ds, & 1 \le i < K, \\ \displaystyle\int_{\tau_{K-1}}^{S} \Phi(\bar{\mathbf{x}}_L; s)\,ds + \mathbf{r}_S, & i = K. \end{cases} \tag{21}$$

Thus,

$$\sum_{i=1}^{K} \mathbf{F}_i = \bar{\mathbf{x}}_L. \tag{22}$$

Each component retains the spatial resolution of the LR input. We use $K = 5$ by default and vary K in the component-count ablation.

These components are separately encoded and combined with the full LR latent in HFRM. The reconstruction identity holds before VAE encoding; it does not imply additive reconstruction in the nonlinear VAE latent space.

Numerical approximation. Let $s_n = n\Delta s$, where Δs is the TV-flow step size. An implicit Euler discretization approximates the flow by

$$u^{n+1} = \arg\min_v \left\{ \frac{1}{2\Delta s}\|v - u^n\|_2^2 + J_{\mathrm{TV}}^h(v) \right\}, \qquad u^0 = f, \tag{23}$$

where J_{TV}^h denotes the discrete TV functional. For interior time samples, the spectral representation can be approximated by

$$\phi_f^n \approx s_n\,\frac{u^{n+1} - 2u^n + u^{n-1}}{(\Delta s)^2}. \tag{24}$$

Component integrals are then approximated by numerical quadrature over their respective scale intervals.

A.3 The effectiveness of fine-tuning MLLM

As shown in Table 6, fine-tuning the MLLM consistently improves downstream SR performance. Compared with the original MLLM, our fine-tuned model improves PSNR from 24.5253 to 25.3913 and SSIM from 0.7157 to 0.7721, while reducing LPIPS and DISTS. More importantly, OCR-A and NED increase from 41.2% and 0.6275 to 48.8% and 0.6921, respectively, demonstrating that degradation-aware fine-tuning provides more reliable textual guidance from LR inputs. Using prompts extracted directly from GT images yields only marginal gains on RealSR, indicating that the fine-tuned MLLM has already recovered most useful semantic and fine-grained information. The remaining gap on UZ-ST suggests that heavily degraded text is still more challenging to infer.

Table 6: Effect of MLLM fine-tuning on downstream super-resolution.

MLLM training	RealSR					UZ-ST	
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑	OCR-A↑ (%)	NED↑
w/o fine-tuning	24.5253	0.7157	0.2698	0.1972	70.9506	41.2	0.6275
w/ fine-tuning (Ours)	25.3913	0.7721	0.2516	0.1817	71.4382	48.8	0.6921
Prompts from GT	25.4013	0.7636	0.2497	0.1849	71.1987	53.1	0.7194

A.4 The Ablation of Three Core Components

In this section, We further investigate the individual contributions and complementarity of the three core components: TDFE, RFPG, and HFRM. As shown in Table 7, the three modules exhibit different but complementary effects. When used individually, TDFE achieves better LPIPS and DISTS than HFRM, demonstrating its effectiveness in preserving fine-grained visual details, while HFRM obtains a relatively higher SSIM, showing its advantage in structural reconstruction. In contrast, RFPG provides the strongest improvement in perceptual and text-related metrics, achieving a CLIP-IQA of 0.6839 and an OCR-A of 47.8%, substantially outperforming TDFE and HFRM alone on UZ-ST. This indicates that fine-grained textual guidance is particularly important for recovering degraded character content.

Combining two components further improves the performance. In particular, TDFE+HFRM achieves the highest PSNR of 25.4194 dB, confirming the complementary effects of explicit detail injection and frequency-aware velocity refinement on reconstruction fidelity. Meanwhile, combinations involving RFPG consistently achieve much higher OCR-A and NED, further demonstrating that RFPG is the main contributor to text reconstruction. Finally, the complete TAPD-SR achieves the best SSIM, LPIPS, DISTS, MUSIQ, CLIPIQA, and NED, while maintaining a competitive PSNR and the highest OCR-A of 48.8%. These results demonstrate that TDFE, RFPG, and HFRM complement each other in detail preservation, fine-grained semantic guidance, and structural refinement, respectively, leading to a better overall fidelity–perception balance.

Table 7: Ablation study of the three core components of TAPD-SR on RealSR and UZ-ST. The best results are highlighted in **bold**.

Components			RealSR						UZ-ST	
TDFE	RFPG	HFRM	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑	CLIPIQA↑	OCR-A↑ (%)	NED↑
✓			25.1719	0.7337	0.2794	0.1998	70.0364	0.6636	35.1	0.5943
	✓		24.8166	0.7197	0.2912	0.2096	70.1715	0.6839	47.8	0.6796
		✓	25.0888	0.7359	0.2837	0.2119	69.5098	0.6593	34.6	0.5924
✓	✓		25.1472	0.7395	0.2696	0.1915	70.9481	0.7034	48.8	0.6893
✓		✓	**25.4194**	0.7557	0.2621	0.1938	70.2864	0.6918	35.1	0.5971
	✓	✓	25.0641	0.7437	0.2689	0.1996	70.4641	0.6981	48.3	0.6874
✓	✓	✓	25.3913	**0.7721**	**0.2516**	**0.1817**	**71.4382**	**0.7258**	**48.8**	**0.6921**

A.5 The superiority analysis of TV spectral decomposition

Our method is to directly perform frequency decomposition from LR input, with the requirement of decomposing without changing the resolution. There are actually several methods currently available to achieve this goal, such as Fourier band decomposition (Fourier Bands) and isotropic undecimated wavelet (IUWT) Starck et al. (2007). In this section, we conduct ablation studies on different decomposition methods and then discuss the superiority of our method.

As shown in Table 8, IUWT consistently outperforms Fourier band decomposition on all metrics, indicating that localized multi-scale representations are more suitable for HFRM than fixed global frequency bands. TV spectral decomposition further achieves the best overall performance. These results suggest that conventional fixed-frequency or predefined multi-scale decomposition is less effective for modeling complex real-world degradations. In contrast, TV spectral decomposition organizes image structures according to nonlinear scales jointly determined by spatial extent and

contrast of image structures. Therefore, TV spectral decomposition enables HFRM to organize LR information according to structure-dependent scales rather than frequency alone, providing more effective multi-scale cues for velocity refinement.

Table 8: Comparison of different resolution-preserving decomposition strategies in HFRM on RealSR. All variants generate five full-resolution components and share the same HFRM architecture and training settings. The best results are highlighted in **bold**.

Decomposition method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	MUSIQ↑	CLIPIQA↑
Fourier Bands	25.2864	0.7506	0.2617	0.1897	71.0273	0.7098
IUWT	25.3472	0.7564	0.2565	0.1876	71.1126	0.7147
Ours	**25.3913**	**0.7721**	**0.2516**	**0.1817**	**71.4382**	**0.7258**

A.6 Model complexity

We further compare the model complexity and inference efficiency of TAPD-SR with other SOTA diffusion-based Real-ISR methods. As shown in Table 9, one-step methods such as TSD-SR and ODTSR achieve substantially lower inference latency due to their single-step generation, requiring only 0.1066 s and 0.828 s, respectively. Multi-step methods generally incur higher computational costs as the number of sampling steps increases. For example, SeeSR and DiT4SR require 2.817 s and 3.3491 s with 50 and 40 sampling steps, respectively.

TAPD-SR adopts the large-scale Qwen-Image backbone and retains 20-step iterative sampling to sufficiently exploit its strong generative prior, resulting in an inference time of 8.512 s. Although this introduces higher computational cost than existing lightweight or distilled methods, TAPD-SR contains 20.51B parameters, which is still lower than the 28.65B parameters of the Qwen-Image-based ODTSR. Overall, our current design prioritizes reconstruction fidelity, perceptual quality, and fine-detail preservation rather than inference efficiency.

Table 9: Model complexity comparison with representative diffusion-based Real-ISR methods. Inference time is measured under the same hardware setting for 128×128 LR inputs on an NVIDIA A6000 GPU.

Method	SeeSR	FaithDiff	TSD-SR	VOSR	DiT4SR	ODTSR	**Ours**
Total Params (B)	2.28	2.70	2.21	1.74	2.72	28.65	20.51
Steps	50	20	1	25	40	1	20
Inference Time (s)	2.817	1.347	0.1066	1.556	3.3491	0.828	8.512

A.7 Limitations and future work

Despite achieving superior fidelity and perceptual quality, TAPD-SR has several limitations. First, the large-parameter diffusion backbone and 20-step iterative sampling incur notable computational complexity and inference latency. Second, under extreme degradations with severe information loss, the MLLM may struggle with character inference and introduce hallucinated descriptions.

In future work, we will explore model distillation and new training methods to develop efficient one-step inference. This can reduce inference latency and model complexity.