PRISM: Prior Rectification and Uncertainty-Aware Structure Modeling for Diffusion-Based Text Image Super-Resolution

Zihang Xu[1]*, **Xiaoyang Liu**[1]*, **Zheng Chen**[1], **Yulun Zhang**[1]†, **Xiaokang Yang**[1]
[1]Shanghai Jiao Tong University

Abstract

Text image super-resolution (Text-SR) requires more than visually plausible detail synthesis: slight errors in stroke topology may alter character identity and break readability. Existing methods improve text fidelity with stronger recognition-based or generative priors, yet they still face two unresolved challenges under severe degradation: the text condition extracted from low-quality inputs can itself be unreliable, and a plausible global prior does not fully determine fine-grained stroke boundaries. We propose **PRISM**, a single-step diffusion-based Text-SR framework that addresses these two challenges through Flow-Matching Prior Rectification (**FMPR**) and a Structure-guided Uncertainty-aware Residual Encoder (**SURE**). FMPR constructs a privileged training-time prior from paired low-quality/high-quality latents and learns a flow matching that transports degraded embeddings toward this restoration-oriented prior space, yielding more accurate and reliable global text guidance. SURE further predicts uncertainty-aware structural residuals to selectively absorb reliable local boundary evidence while suppressing ambiguous stroke cues. Together, these components enable explicit global prior rectification and local structure refinement within a single diffusion restoration pass. Experiments on both synthetic and real-world benchmarks show that PRISM achieves state-of-the-art performance with millisecond-level inference. Our dataset and code will be available at `https://github.com/faithxuz/PRISM`.

1 Introduction

Text image super-resolution (Text-SR) aims to restore high-resolution text images from degraded low-resolution inputs. Unlike generic image super-resolution [1–3], text is both visual and symbolic. A small artifact in a natural texture may only affect perceptual quality, whereas a broken stroke, merged component, or distorted enclosure can change the identity of a character. This sensitivity is especially severe for densely structured scripts such as Chinese [4], where subtle stroke layouts often distinguish different characters. An effective Text-SR system must therefore recover not only visually plausible details, but also semantically faithful glyph structures with sub-character precision.

Existing Text-SR methods address this structure-sensitive problem by introducing stronger text-specific guidance. Early methods [5–7] improve readability with recognition supervision, sequential modeling, and layout-aware reasoning. These cues help the model reason about text, but can become unreliable when severe degradation removes stroke evidence needed for character discrimination. Later methods [4, 8] introduce richer text-specific priors, such as generative character-structure priors and text style embeddings, to handle complex glyphs and appearance variation. Recent diffusion-based Text-SR and text-aware restoration methods [9–12] further exploit generative priors, text diffusion, segmentation, or text-spotting guidance to improve perceptual realism and text fidelity. While these developments highlight the importance of text-aware guidance, its reliability under severe degradation and its effective translation into local stroke geometry remain insufficiently addressed.

*Equal contribution
†Corresponding author: Yulun Zhang, yulun100@gmail.com

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

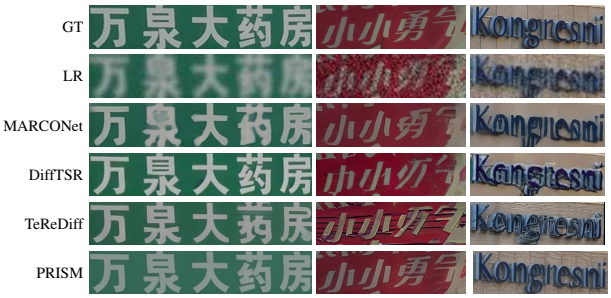

Figure 1: Entangled objectives cause structural errors.

Instead of debating whether to incorporate text-aware cues, the current bottleneck lies in how to obtain them reliably under severe degradation. In recent diffusion-based methods, text conditions are typically derived directly from the degraded input. When strokes are heavily corrupted, these inferred conditions are inherently unreliable. Because condition estimation and image reconstruction are entangled under a shared objective, the model cannot distinguish between correcting stroke geometry and compensating for an erroneous high-level condition, often yielding sharp but semantically incorrect outputs (Fig. 1). Moreover, even if a plausible global semantic condition is obtained, it cannot fully determine pixel-aligned local structures such as stroke closures and intersections. Directly relying on edge cues from the degraded image to fill this gap is equally risky, as the visible edges are often missing or misleading. These coupled challenges suggest a need for an explicit decomposition: we could first recover a stable text-aware latent condition from the degraded input, and subsequently refine uncertain local stroke geometry in image space under that guidance.

We propose **PRISM**, a single-step Text-SR framework based on pre-trained Diffusion Models (DMs), with **P**rior **R**ectification and uncerta**I**nty-aware **S**tructure **M**odeling. PRISM explicitly decomposes restoration into global prior rectification and local structure refinement. Its first component, **FMPR** (**F**low-**M**atching **P**rior **R**ectification), constructs a privileged training-time prior from paired LQ/HQ latents and learns a flow matching that transports the LQ embedding distribution toward this privileged prior space. Unlike conventional diffusion-style prior extraction that starts from pure noise or treats the inferred prior as a static side condition, FMPR directly models the velocity field from degraded embeddings to restoration-oriented text tokens, producing more accurate and reliable global guidance.

The second component, **SURE** (**S**tructure-guided **U**ncertainty-aware **R**esidual **E**ncoder), injects residual controls to refine local stroke geometry. SURE is a structure-aware encoder branch that predicts both the mean and uncertainty of structural features, allowing the model to selectively absorb reliable boundaries while suppressing ambiguous ones, instead of treating LQ edges as deterministic truth. This uncertainty-aware design is particularly important for Text-SR, where an overconfident wrong edge can be more harmful than a missing edge. To the best of our knowledge, this is the first uncertainty-aware boundary control formulation tailored to text-specific structural refinement.

PRISM keeps the efficiency advantage of one-step restoration while substantially improving the quality of text-aware guidance and structure recovery. The FMPR flow transport is performed in a compact embedding space, and the final image restoration still uses a single diffusion backbone call, making the overall system significantly faster than iterative diffusion-based Text-SR while preserving superior generative quality. Experiments on both synthetic and real-world benchmarks show that PRISM achieves state-of-the-art overall performance with millisecond-level inference.

Our contributions are summarized as follows:

- We revisit Text-SR from the perspective of *prior reliability* and *structural uncertainty*, and propose **PRISM**, a Text-SR model with single-step diffusion inference.
- We propose **FMPR**, a flow-matching prior rectification module that learns to transport LQ text embeddings toward a privileged HQ-aware prior space and injects the recovered tokens into the main backbone for efficient restoration.
- We propose **SURE**, an uncertainty-aware structure guidance module that predicts stochastic edge features and adaptively gates boundary information through uncertainty learning, yielding more robust local structure control under severe degradation.
- Extensive experiments on both synthetic and real-world benchmarks show that PRISM achieves state-of-the-art performance at the millisecond level.

2 Related Works

Real-World Image Super-Resolution. Real-world image super-resolution (Real-SR) aims to restore high-quality images from low-resolution inputs with complex and unknown degradations. Early methods mainly improve robustness through degradation modeling and discriminative reconstruction,

such as BSRGAN [13] and Real-ESRGAN [14]. With the development of generative models [15], recent methods exploit diffusion priors to recover realistic details under severe degradation [16–23]. For example, DiffBIR [17] decomposes blind restoration into degradation removal and diffusion-based detail regeneration, while SUPIR [21] scales generative restoration with large diffusion priors and high-quality data. Since iterative diffusion sampling is expensive, efficient Real-SR methods further compress or reformulate diffusion restoration into few-step or one-step inference [1–3, 24–26]. OSEDiff [1], for instance, performs one-step Real-SR by directly starting from the low-quality image. Stronger generative backbones, including SDXL [27], DiT [28], SD3 [29], and FLUX [30], have also been studied or adapted for restoration [31, 3]. However, these methods mainly target generic natural image restoration and lack dedicated modeling for character identity and stroke structure.

Text Image Super-Resolution. Text image super-resolution (Text-SR) focuses on restoring readable text crops or text-line images from degraded inputs. Different from generic SR, Text-SR requires the restored image to preserve character identity as well as visual quality. Early methods address this problem by introducing recognition guidance, sequential reasoning, layout modeling, and text-prior attention [5–7, 32, 33]. TSRN [5] frames Text-SR as a recognition-oriented restoration problem, while TBSRN [6] and TATT [7] further exploit text layouts, character details, and deformation-aware attention. Later studies move from high-level recognition cues toward more explicit text structure modeling [4, 8, 34–39]. These works shift the focus from recognizing text to preserving how characters are spatially organized and visually presented. MARCONet [4] learns a generative structure prior for blind text restoration, while StyleSRN [8] complements text priors with style embeddings to better preserve appearance details. More recently, diffusion-driven Text-SR methods have explored generative restoration under text-specific conditions [40, 9]. DiffTSR [9] couples image and text diffusion, demonstrating the potential of diffusion priors for severely degraded Text-SR.

A closely related direction studies text-aware restoration in broader real-world or full-image settings [10–12]. These methods usually build upon general restoration frameworks and introduce text awareness through text-region perception, segmentation, text spotting, or text-aware conditioning. TADiSR [10] integrates text-aware attention and joint segmentation decoders for real-world image SR, while TeReDiff [11] couples diffusion restoration with a text-spotting module. Although these works operate on full images, their text-related component is closely connected to crop-level Text-SR: full-image text-aware restoration still requires reliable restoration of local text regions, while crop-level Text-SR isolates this text-centric subproblem and enables more focused modeling of character fidelity and stroke structures. Thus, the two settings are mutually convertible and complementary. Following this rationale, we adopt the crop-level setting and focus on text-line super-resolution. By isolating the problem at the crop level, we are able to design highly dedicated modules for reliable text prior recovery and uncertainty-aware stroke refinement. Furthermore, our method can be seamlessly integrated into full-image restoration pipelines as a robust, dedicated text-enhancing module.

3 Methodology

3.1 Overall Structure

The overall structure of our **PRISM** is illustrated in Fig. 2. Built upon a pre-trained latent diffusion model [15], our method follows a progressive restoration paradigm for Text-SR. Severe text degradation introduces two coupled challenges: the text-aware condition inferred from the degraded input may be unreliable, while fine-grained stroke topology and boundary placement may remain ambiguous even with a plausible prior. To address these, we first learn a recoverable text prior and then refine spatially unstable structures under the recovered prior.

Given a degraded text image x_l, the frozen VAE encoder maps it to a latent representation $z_l = \mathcal{E}_{\text{vae}}(x_l)$. The prior recovery branch **FMPR** predicts a text-aware embedding $\hat{c}$ from z_l, where $\hat{c}$ is learned to approximate a privileged prior space constructed from paired training data. In parallel, the structure control branch **SURE** extracts uncertainty-aware spatial cues from x_l and predicts multi-level residual controls $\mathcal{R} = \{r_i\}_{i=1}^{M}$. Following [1, 41], the single-step restoration is computed as $\hat{z}_h = \frac{z_l - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}}$ [42, 15], where z_l is the degraded latent at a fixed timestep t, $\bar{\alpha}_t$ is the noise schedule coefficient, and $\hat{\epsilon}$ is the predicted noise. For brevity, we denote the overall process as:

$$\hat{z}_h = \mathcal{U}_{\bar{\theta}}\left(z_l, \hat{c}; \mathcal{R}\right), \qquad \hat{x} = \mathcal{D}_{\text{vae}}(\hat{z}_h), \tag{1}$$

where $\mathcal{U}_{\bar{\theta}}$ denotes the diffusion backbone used in the final stage, and $\mathcal{D}_{\text{vae}}$ is the VAE decoder. For clarity, we use θ_{p}, θ_{r}, and $\bar{\theta}$ to denote the restoration backbone after privileged-prior construction, after recoverable-prior learning, and after training for structure control, respectively.

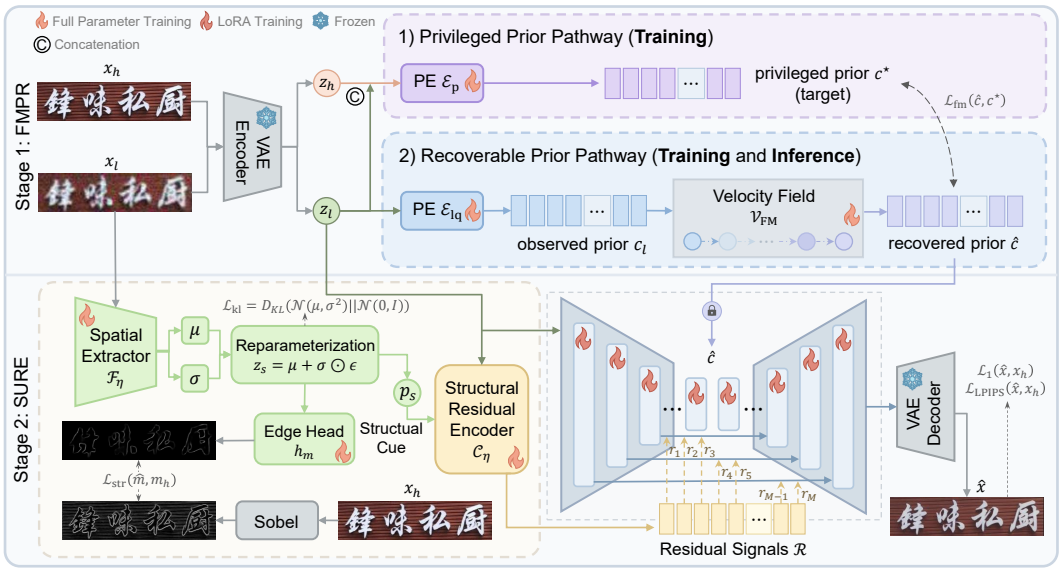

Figure 2: Overall structure of our PRISM.

During training, we first construct a privileged conditional prior from paired LQ/HQ latents and learn to recover it from the degraded input alone. After the recoverable prior pathway is trained, we freeze both the prior pathway and the restoration backbone and optimize the structure control branch. During inference, the model only requires the degraded input x_l: the prior branch produces $\hat{c}$, the structure branch produces $\mathcal{R}$, and the restoration backbone generates the final output.

3.2 FMPR: Flow-Matching Prior Rectification

A reliable text-aware condition is crucial for Text-SR but difficult to obtain under severe degradation. Direct extraction from degraded images often yields unreliable priors that misguide restoration. Thus, our goal is not merely to apply a text prior, but to learn one that is informative during training and recoverable from degraded observations at test time.

Our solution, FMPR, decouples prior construction from prior recovery. During training, paired high-quality and low-quality data allow us to construct a privileged conditional prior that defines a target prior space. At inference, where only degraded inputs are available, we learn an LQ-only recovery path to map observations toward this privileged space. This follows the spirit of learning with privileged information [43–45]: extra information available only during training defines a more reliable learning target, while the inference model remains dependent solely on observed inputs.

Privileged Conditional Prior. Given a paired training sample (x_l, x_h), we encode both images into the latent space as $z_l = \mathcal{E}_{\text{vae}}(x_l)$ and $z_h = \mathcal{E}_{\text{vae}}(x_h)$ with the frozen VAE encoder. A prior encoder (PE) $\mathcal{E}_{\text{p}}$ takes the concatenated LQ-HQ latents and produces a privileged conditional prior. The privileged-prior construction is formulated as

$$c^\star = \mathcal{E}_{\text{p}}([z_l; z_h]), \qquad c^\star \in \mathbb{R}^{N\times D}, \tag{2}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation and N, D are the token number and channel dimension. Since $c^\star$ sees both degraded evidence and target latent structure, it provides a cleaner conditional signal than an LQ-only prior. We use it as the text embedding to warm up the one-step backbone, where $c^\star$ serves as the key (K) and value (V) for the UNet cross-attention layers:

$$\hat{z}_h^\star = \mathcal{U}_{\theta_{\text{p}}}(z_l, c^\star), \qquad \hat{x}^\star = \mathcal{D}_{\text{vae}}(\hat{z}_h^\star), \qquad \mathcal{L}_{\text{priv}} = \|\hat{x}^\star - x_h\|_1 + \lambda_{\text{lpips}}\mathcal{L}_{\text{LPIPS}}(\hat{x}^\star, x_h). \tag{3}$$

Importantly, $c^\star$ is only available during training; it defines the target prior distribution rather than a test-time condition. Its role is to define a privileged prior space that specifies what an informative text-aware condition should look like for restoration.

Recoverable Prior Learning. After the privileged prior space is established, the remaining problem is how to approximate it without access to x_h. We first map the degraded latent to an observed prior $c_l = \mathcal{E}_{\text{lq}}(z_l)$ using an LQ-only PE $\mathcal{E}_{\text{lq}}$ with the same structure as $\mathcal{E}_{\text{p}}$.

A straightforward alternative is to directly regress $c^\star$ from c_l. However, under severe degradation, the mapping from the observed prior to the privileged prior can be highly ambiguous. Motivated by flow-matching generative modeling [46, 47], we formulate prior recovery as a flow-matching transport problem.

Specifically, we learn a velocity field $\mathcal{V}_{\mathrm{FM}}$ over the conditional embedding space and integrate it from the observed prior. For each paired sample $(c_l, c^\star)$, we define the straight interpolation path and its corresponding velocity as reference:

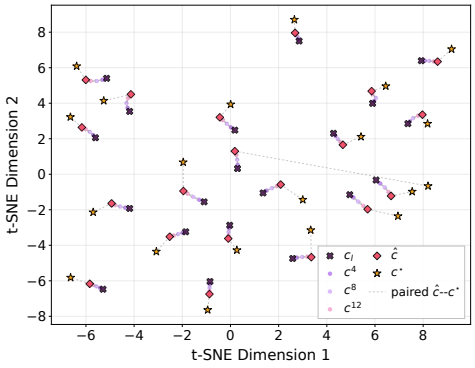

Figure 3: FMPR prior recovery trajectory.

$$c_{\mathrm{ref}}(t) = (1-t)c_l + tc^\star, \quad u_{\mathrm{ref}}(t) = \frac{dc_{\mathrm{ref}}(t)}{dt} = c^\star - c_l. \tag{4}$$

Because the latent space is highly compact, we integrate $\mathcal{V}_{\mathrm{FM}}$ which governs the actual transport trajectory using K Euler steps for both training and inference. Specifically, we apply:

$$c^{k+1} = c^k + \frac{1}{K}\mathcal{V}_{\mathrm{FM}}\left(c^k, \frac{k}{K}\right), \tag{5}$$

starting from $c^0 = c_l$ and obtaining the recovered prior $\hat{c} = c^K$, as visualized for 20 representative samples in Fig. 3, where c_l, $\hat{c}$, $c^\star$ and intermediate states $\{c^4, c^8, c^{12}\}$ are projected into a 2D t-SNE space. Then, $\hat{c}$ is used as the text-aware condition for restoration: $\hat{z}_h^{\mathrm{r}} = \mathcal{U}_{\theta_{\mathrm{r}}}(z_l, \hat{c}), \hat{x}^{\mathrm{r}} = \mathcal{D}_{\mathrm{vae}}(\hat{z}_h^{\mathrm{r}})$, where $\mathcal{U}_{\theta_{\mathrm{r}}}$ is initialized from the privileged-prior backbone $\mathcal{U}_{\theta_{\mathrm{p}}}$ and further adapted with the recovered prior. The objective combines image-level restoration supervision and latent prior matching:

$$\mathcal{L}_{\mathrm{stage1}} = \underbrace{\|\hat{x}^{\mathrm{r}} - x_h\|_1 + \lambda_{\mathrm{lpips}}\mathcal{L}_{\mathrm{LPIPS}}(\hat{x}^{\mathrm{r}}, x_h)}_{\mathcal{L}_{\mathrm{img}}} + \lambda_{\mathrm{fm}} \underbrace{\|\hat{c} - c^\star\|_1}_{\mathcal{L}_{\mathrm{fm}}}. \tag{6}$$

This stage stabilizes the high-level text-aware condition under severe degradation, guiding the model toward plausible character identities and coarse structures. However, the recovered prior is still an embedding-space condition, which does not explicitly determine where uncertain local stroke boundaries should be placed in the image. This motivates the next stage, which performs explicit structure refinement under the recovered prior.

3.3 SURE: Structure-guided Uncertainty-aware Residual Encoder

FMPR learned in Sec. 3.2 stabilizes global text identity, but local stroke boundaries can still be ambiguous. To address this, after the recoverable prior is learned, we freeze the recovered-prior pathway and the backbone, and train a structure-guided uncertainty-aware residual encoder (SURE). Specifically, SURE consists of two cascading modules: an uncertainty-aware spatial cue extractor $\mathcal{F}_\eta$ and a structural residual encoder $\mathcal{C}_\eta$. SURE focuses exclusively on local structural correction.

Uncertainty-Aware Spatial Cue Extraction. The degraded input contains partial but unevenly reliable structural evidence. Since LQ-derived edges may be incomplete or misleading, treating them as deterministic constraints can amplify degradation artifacts or hallucinate incorrect boundaries. We therefore model the spatial cue in an uncertainty-aware manner, following the general practice of uncertainty-aware prediction for ambiguous visual evidence [48–50].

A spatial cue extractor $\mathcal{F}_\eta$ first produces a feature map $f = \mathcal{F}_\eta(x_l)$. From f, two lightweight heads predict the mean and log-variance of a latent structural cue distribution, denoted as $\mu = h_\mu(f)$ and $\log\sigma^2 = h_\sigma(f)$. We then sample a stochastic structural cue via reparameterization:

$$z_s = \mu + \sigma \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad \sigma = \exp\left(\frac{1}{2}\log\sigma^2\right). \tag{7}$$

Compared with a deterministic cue, this formulation allows ambiguous regions to be represented with higher uncertainty instead of forcing all local evidence into a single confident estimate. The sampled cue z_s is projected into the structure control space as $p_s = \Pi(z_s)$, and simultaneously decoded by an edge head into an auxiliary boundary map $\hat{m} = h_m(z_s)$ for loss regulation.

Structure Control Branch. Let $\hat{c}$ denote the recovered prior in Sec. 3.2. Given the degraded latent z_l, recovered prior $\hat{c}$, and projected structural cue p_s, the structure control branch predicts residual signals $\mathcal{R}$ that are then injected into the skip-connection features of the frozen UNet $\mathcal{U}_{\bar{\theta}}$:

$$\mathcal{R} = \{r_i\}_{i=1}^{M} = \mathcal{C}_\eta\left(z_l, \hat{c}, p_s\right), \qquad \hat{z}_h^{\mathrm{s}} = \mathcal{U}_{\bar{\theta}}\left(z_l, \hat{c}; \mathcal{R}\right), \qquad \hat{x}^{\mathrm{s}} = \mathcal{D}_{\mathrm{vae}}(\hat{z}_h^{\mathrm{s}}), \tag{8}$$

where $\mathcal{C}_\eta$ is the structural residual encoder and is encouraged to improve restoration through spatial refinement rather than by re-estimating the text-aware condition.

In practice, we implement $\mathcal{C}_\eta$ by initializing its architecture and weights from the diffusion backbone's encoder for simplicity. This allows image-space structural cues to be injected into multiple layers of the frozen backbone while preserving the prior-guided capability learned in the previous stage.

Training objective. To ensure that the structure branch learns meaningful stroke-level refinement rather than arbitrary feature perturbation, we impose explicit structure-aware supervision. We use the Sobel operator to extract a boundary target $m_h = \mathcal{S}(x_h)$ from the clean image. We further impose a KL penalty between the predicted latent distribution and a standard Gaussian prior. This prevents the variance from collapsing to zero or becoming arbitrarily unstable, thereby preserving the uncertainty-aware nature of the structural cue. The full objective for structure control is:

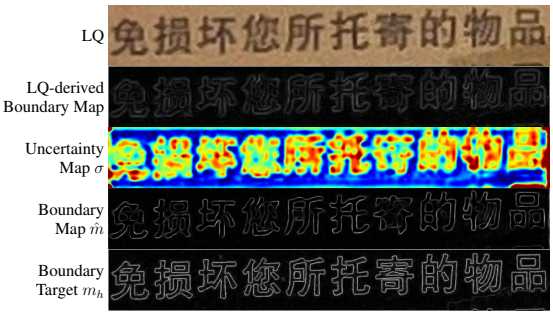

Figure 4: SURE structural cue visualization.

$$\mathcal{L}_{\mathrm{stage2}} = \underbrace{\|\hat{x}^{\mathrm{s}} - x_h\|_1 + \lambda_{\mathrm{lpips}}\mathcal{L}_{\mathrm{LPIPS}}(\hat{x}^{\mathrm{s}}, x_h)}_{\mathcal{L}_{\mathrm{img}}} + \lambda_{\mathrm{str}} \underbrace{\|\hat{m} - m_h\|_1}_{\mathcal{L}_{\mathrm{str}}} + \lambda_{\mathrm{kl}} \underbrace{D_{\mathrm{KL}}\left(\mathcal{N}(\mu, \sigma^2) \,\|\, \mathcal{N}(0, I)\right)}_{\mathcal{L}_{\mathrm{kl}}}. \tag{9}$$

As visualized in Fig. 4 (LQ, LQ-derived boundary map, uncertainty map σ, $\hat{m}$, and m_h), the model generates distinctly clearer structures in $\hat{m}$ where it exhibits high confidence (i.e., low uncertainty, indicated by the red areas in the uncertainty map), whereas regions with high uncertainty appear correspondingly blurry in $\hat{m}$. By feeding these uncertainty-aware regularized features into $\mathcal{C}_\eta$, the model can more effectively focus on local stroke topology, boundary closure, and spatial alignment.

4 Experiments

4.1 Experimental Setup

Datasets. We focus on Chinese-English text-line SR. Existing Text-SR datasets differ in language coverage, image quality, scale, and task scope, making it difficult to form a consistent training corpus for this task. TextZoom [5] provides real-world English pairs but lacks broader bilingual coverage. Real-CE [51] contains Chinese-English real text pairs, but is relatively limited in scale. SA-Text [11] provides high-quality scene images with dense text annotations, but our corpus analysis shows limited usable Chinese text crops, as detailed in the appendix. We therefore construct BTL by combining filtered real text crops from existing annotated sources with synthetic text-line rendering.

Specifically, we collect Chinese text crops with annotations from the CTR benchmark [52], extract English text crops from SA-Text annotations [11], and include digit samples from both sources. All candidates are filtered by unified criteria: (i) valid annotations; (ii) resized height of 128 pixels; (iii) aspect ratios between 2 and 8; (iv) transcripts no longer than 24 characters; and (v) no-reference IQA-based quality ranking. For quality ranking, we use a weighted score based on MUSIQ [53], MANIQA [54], and CLIP-IQA [55]. This process yields 50K quality-controlled real text-line images.

To improve text appearance and layout diversity, we further synthesize 50K high-quality text-line images following the synthetic text rendering strategy of MARCONet [4]. Together with the curated real crops, this forms BTL, a 100K HQ bilingual text-line corpus. For each HQ image, we generate an LQ counterpart using degradation pipelines based on BSRGAN [13] and Real-ESRGAN [14]. We use 80K image pairs for training and reserve 20K pairs for synthetic evaluation, denoted as BTL-train and BTL-test, respectively.

We further evaluate real-world generalization on RealCE-val. Since some LQ-HQ pairs exhibit noticeable misalignment, color mismatch, or annotation errors, we filter invalid pairs and manually

(a) Synthetic dataset BTL-test.

Methods	×2					×4				
	PSNR ↑	LPIPS ↓	FID ↓	ACC ↑	NED ↑	PSNR ↑	LPIPS ↓	FID ↓	ACC ↑	NED ↑
TSRN [5]	22.28	0.3581	48.85	52.63%	0.7363	21.53	0.5039	98.65	31.37%	0.5046
TBSRN [6]	24.47	0.3619	65.74	51.69%	0.7243	<u>22.25</u>	0.4865	95.46	33.41%	0.5241
TATT [7]	24.51	0.3584	57.40	50.07%	0.7154	20.20	0.4978	83.73	29.80%	0.4826
MARCONet [4]	**25.03**	<u>0.1974</u>	17.88	59.08%	0.8075	**22.35**	<u>0.2956</u>	33.48	<u>37.96%</u>	<u>0.6049</u>
DiffTSR [9]	24.21	0.2327	15.96	**59.85%**	<u>0.8117</u>	21.00	0.3352	<u>25.55</u>	37.85%	0.6007
StyleSRN [8]	12.44	0.7272	119.89	26.55%	0.4242	12.30	0.7723	142.33	19.05%	0.3248
TeReDiff [11]	22.38	0.2692	<u>13.22</u>	49.70%	0.7272	20.16	0.3832	27.14	29.14%	0.4910
PRISM	<u>24.53</u>	**0.1514**	**6.14**	<u>59.78%</u>	**0.8220**	22.08	**0.2314**	**12.57**	**42.12%**	**0.6644**

(b) Real-world dataset RealCE-val.

Methods	×2					×4				
	PSNR ↑	LPIPS ↓	FID ↓	ACC ↑	NED ↑	PSNR ↑	LPIPS ↓	FID ↓	ACC ↑	NED ↑
TSRN [5]	19.51	0.1420	58.88	82.64%	0.9392	<u>19.32</u>	0.2988	141.70	52.17%	0.7860
TBSRN [6]	20.32	**0.1343**	51.66	83.99%	**0.9480**	19.04	0.2767	111.07	60.17%	0.8324
TATT [7]	<u>20.54</u>	0.2118	80.67	80.71%	0.9385	18.16	0.3048	122.78	46.67%	0.7538
MARCONet [4]	19.38	0.1704	54.03	**84.36%**	0.9377	18.91	<u>0.2326</u>	<u>74.52</u>	<u>60.62%</u>	<u>0.8381</u>
DiffTSR [9]	20.26	0.1964	<u>35.59</u>	81.97%	0.9424	18.60	0.3052	74.94	54.00%	0.8077
StyleSRN [8]	14.12	0.5009	90.51	51.30%	0.7043	13.97	0.5403	101.39	38.67%	0.6083
TeReDiff [11]	17.81	0.2913	88.15	52.27%	0.6969	17.03	0.3632	106.13	34.23%	0.5675
PRISM	**21.00**	<u>0.1372</u>	**33.71**	<u>84.28%</u>	<u>0.9442</u>	**19.89**	**0.2043**	**47.83**	**65.19%**	**0.8521**

Table 1: Quantitative comparison on BTL-test and RealCE-val under $\times 2$ and $\times 4$ text image super-resolution. Best and second-best results are shown in bold and underlined, respectively.

correct erroneous annotations, resulting in 1,037 valid testing pairs. Details of BTL construction and the RealCE-val evaluation protocol are provided in the appendix.

Implementation Details. We build our model on the pretrained Stable Diffusion 2.1-base model and train the UNet with LoRA[56] of rank 16. FMPR contains two training stages, privileged-prior construction and LQ-only prior recovery, each trained for 100K iterations. SURE is then trained for 50K iterations with the FMPR pathway and restoration backbone frozen. All stages use AdamW with a learning rate of 5×10^{-5} and a total batch size of 8 on two NVIDIA RTX A6000 GPUs. FMPR uses 16-step Euler discretization in Eq. (5) to recover the text prior, and final restoration is performed in one step at a fixed timestep of 399 [41, 57]. We set $\lambda_{\text{lpips}} = 1$ and $\lambda_{\text{fm}} = 1$ for FMPR, and $\lambda_{\text{lpips}} = 1$, $\lambda_{\text{str}} = 1$, and $\lambda_{\text{kl}} = 0.1$ for SURE.

Compared Methods and Evaluation Metrics. We compare our method with representative Text-SR methods, including TSRN [5], TBSRN [6], TATT [7], MARCONet [4], DiffTSR [9], and StyleSRN [8]. We also include TeReDiff [11], a recent text-aware image restoration method. For fair comparison, all baselines with publicly available training implementations are retrained or fine-tuned on BTL-train following their respective training procedures. StyleSRN [8] is evaluated zero-shot using its released checkpoint as its training implementation is unavailable. We evaluate reconstruction fidelity with peak signal-to-noise ratio (PSNR) and learned perceptual image patch similarity (LPIPS) [58], which measure image-space and feature-space differences from the reference image, respectively. We use Fréchet inception distance (FID) [59] to assess distributional realism. For text fidelity, we report OCR accuracy (ACC) and normalized edit distance (NED) [51], both computed using PP-OCRv5 [60] as the recognition model.

4.2 Main Results

Quantitative Comparisons. Table 1 reports quantitative comparisons on BTL-test and RealCE-val under $\times 2$ and $\times 4$ settings. On BTL-test, our method obtains the best LPIPS, FID, and NED at both scales, and the highest ACC under $\times 4$. These results show clear advantages in perceptual quality and text fidelity. Although there is a gap in PSNR, it aligns with the well-known perception-distortion tradeoff: unlike PSNR-oriented methods that tend to produce overly smoothed outputs, our diffusion-based approach recovers sharp, high-frequency stroke details that significantly benefit character readability. While our model is trained on BTL-train, it also generalizes well to real-world degraded text images. On RealCE-val, our method achieves the best PSNR and FID and ranks second in the remaining metrics under $\times 2$, and ranks first across all metrics under the more challenging $\times 4$ setting. Notably, under $\times 4$, it improves ACC from 60.62% to 65.19% and reduces FID from 74.52 to 47.83



Figure 5: Qualitative comparison on the synthetic BTL-test dataset for $\times 4$ super-resolution. We compare our method with TATT [7], StyleSRN [8], MARCONet [4], DiffTSR [9], and TeReDiff [11].

Methods	PSNR ↑	LPIPS ↓	DISTS ↓	FID ↓	ACC ↑	NED ↑
Base model	19.8113	0.2121	0.2143	53.42	62.39%	0.8390
Privileged reference ($c^\star$)	20.7509	0.1211	0.1604	31.49	64.90%	0.8588
Direct regression	19.8167	0.2107	0.2123	50.67	64.03%	0.8446
Diffusion	**19.9202**	0.2123	0.2109	49.40	62.49%	0.8371
FMPR	19.8257	**0.2059**	**0.2097**	**47.54**	**64.61%**	**0.8461**

Table 2: Ablation study of prior learning paradigms on RealCE-val.

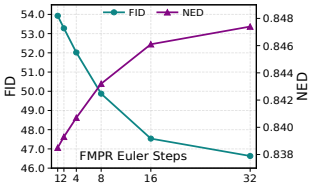

Figure 6: FMPR Euler steps.

compared with the second-best results. Overall, the results show that the proposed method improves perceptual realism and character-level readability, especially under severe real-world degradation.

Qualitative Comparisons. Figures 5 and 7 compare visual results on BTL-test and RealCE-val. As can be seen, TATT [7] and StyleSRN [8] tend to produce over-smoothed text, especially for complex Chinese glyphs under severe blur. MARCONet [4] restores sharper strokes in some cases, but often introduces structural distortion or weak text-background consistency, as shown in the 2nd and 3rd BTL-test examples. Diffusion-based methods improve perceptual sharpness, but still suffer from text-specific artifacts. DiffTSR [9] produces broken or merged strokes under severe degradation, as shown in the 4th BTL-test and 2nd RealCE-val examples. TeReDiff [11] is prone to false-color artifacts on small text images and may hallucinate redundant or incorrect strokes, as shown in the 4th BTL-test and 3rd RealCE-val examples. In contrast, by combining recovered text priors with uncertainty-aware structural cues, our method better preserves character readability and local stroke continuity while maintaining more consistent background appearance.

Inference Efficiency. Inference efficiency is important for practical Text-SR, especially for diffusion models. Our PRISM uses only one denoising step, compared with 200 steps for DiffTSR [9] and 50 for TeReDiff [11]. We compare speed on test images resized to 128×512. For a single image, our method takes 0.08 s, compared with 10.70 s for DiffTSR and 5.27 s for TeReDiff. Importantly, this single-step design makes our inference speed highly comparable to standard CNN- and Transformer-based methods. Detailed runtime comparisons of all evaluated methods are provided in the appendix.

4.3 Ablation Studies

Analysis of Prior Learning Paradigms. We compare different paradigms for recovering the privileged prior from degraded observations on RealCE-val. As shown in Tab. 2, using the privileged condition $c^\star$ provides an HQ-aware reference, confirming that paired LQ/HQ latents define an informative text-aware prior space. The remaining variants examine how such a prior can be approximated from the degraded input alone. Direct regression yields moderate gains but struggles to close the prior gap, as it treats recovery merely as target fitting, which is insufficient to close the gap between degraded and privileged priors. While the diffusion-based variant improves reconstruction, its gains in recognition-oriented metrics remain limited. This suggests that the traditional diffusion approach, constructing the prior from a pure Gaussian distribution under the LQ condition, entails



Figure 7: Qualitative comparison on the real-world RealCE-val dataset for $\times 4$ super-resolution. We compare our method with TATT [7], StyleSRN [8], MARCONet [4], DiffTSR [9], and TeReDiff [11].

Methods	PSNR ↑	SSIM ↑	LPIPS ↓	ACC ↑	NED ↑
FMPR only	19.8257	0.5968	0.2059	64.61%	0.8461
Residual branch only	19.8671	0.5947	0.2099	64.80%	0.8468
w/o uncertainty	19.8768	0.5974	0.2056	64.80%	0.8477
Full model	**19.8919**	**0.6012**	**0.2043**	**65.19%**	**0.8521**

Table 3: Ablation study of SURE on RealCE-val.



Figure 8: SURE visual details.

overly complex and redundant generation paths. This unnecessary complexity hinders the effective learning of strict character structures. In contrast, FMPR starts from the observed prior and learns a continuous transport field toward the privileged prior space. This effectively rectifies unreliable information while preserving the degraded condition, achieving the best overall balance of character-level fidelity and perceptual quality. We further compare FMPR against a compute-matched recurrent refinement variant. Results across three independent seeds consistently favor FMPR (in the appendix).

Analysis of FMPR Euler Steps. We further study the effect of FMPR Euler steps K. As shown in Fig. 6, increasing K consistently improves perceptual realism and text fidelity, indicating that one-step rectification is insufficient for reliable prior recovery. The gains gradually saturate as more steps are used, suggesting that FMPR quickly approaches a stable region in the prior space. While $K = 32$ yields slight improvements, it doubles the computational cost compared with $K = 16$. We therefore choose $K = 16$ as a practical trade-off between restoration performance and efficiency.

Analysis of SURE. We ablate SURE on RealCE-val by progressively removing its uncertainty and structural guidance. As shown in Tab. 3, a plain residual branch without any edge information input brings limited benefit and even weakens perceptual reconstruction. Introducing deterministic edge guidance improves performance by providing complementary stroke-level information. However, deterministic guidance remains vulnerable to severe degradation, where unreliable stroke evidence may be injected under confident conditions and thus limits character-level recovery. The full model addresses this issue with uncertainty learning, allowing the structural branch to model ambiguous regions instead of enforcing a deterministic prediction. This consistently improves restoration quality and text fidelity. The zoomed-in patches in Fig. 8 further illustrate this effect. Compared with the base model and FMPR-only result, the full model produces cleaner local structures and more stable stroke topology. For the first column, it restores the letter "e" more accurately. And for the last three Chinese characters, the full model preserves more accurate strokes and tighter structural closure. Results across three independent seeds support the stability of these gains, as detailed in the appendix.

5 Conclusion

We propose **PRISM**, a single-step diffusion-based framework for Text-SR that addresses two coupled ambiguities under severe degradation: unreliable text-aware prior estimation and uncertain local stroke structures. PRISM decomposes the restoration process into global prior rectification and local

structure refinement. FMPR constructs a privileged prior space from paired LQ/HQ latents and learns to recover a reliable text-aware condition from degraded inputs through flow matching. SURE further injects uncertainty-aware structural residuals into the frozen restoration backbone, allowing the model to refine ambiguous stroke boundaries without over-committing to unreliable LQ edge evidence. This design preserves the efficiency of one-step diffusion restoration while improving both character fidelity and perceptual quality. Experiments on synthetic and real-world benchmarks demonstrate that PRISM achieves superior performance over representative Text-SR and text-aware restoration methods, especially under severe degradation and complex glyph structures.

Acknowledgments and Disclosure of Funding

Funding. This work was supported by the CAAI-Tencent Rhino-Bird Open Research Fund.

Competing Interests. The authors declare no additional competing financial interests.

References

[1] Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *NeurIPS*, 2024.

[2] 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 *CVPR*, 2025.

[3] Jianze Li, Jiezhang Cao, Yong Guo, Wenbo Li, and Yulun Zhang. One diffusion step to real-world super-resolution via flow trajectory distillation. In *ICML*, 2025.

[4] Xiaoming Li, Wangmeng Zuo, and Chen Change Loy. Learning generative structure prior for blind text image super-resolution. In *CVPR*, 2023.

[5] Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *ECCV*, 2020.

[6] Jingye Chen, Bin Li, and Xiangyang Xue. Scene text telescope: Text-focused scene image super-resolution. In *CVPR*, 2021.

[7] Jianqi Ma, Zhetong Liang, and Lei Zhang. A text attention network for spatial deformation robust scene text image super-resolution. In *CVPR*, 2022.

[8] Shengrong Yuan, Runmin Wang, Ke Hao, Xuqi Ma, Changxin Gao, Li Liu, and Nong Sang. Stylesrn: Scene text image super-resolution with text style embedding. In *ICCV*, 2025.

[9] Yuzhe Zhang, Jiawei Zhang, Hao Li, Zhouxia Wang, Luwei Hou, Dongqing Zou, and Liheng Bian. Diffusion-based blind text image super-resolution. In *CVPR*, 2024.

[10] Qiming Hu, Linlong Fan, Yiyan Luo, Yuhang Yu, Xiaojie Guo, and Qingnan Fan. Text-aware real-world image super-resolution via diffusion model with joint segmentation decoders. In *NeurIPS*, 2025.

[11] Jaewon Min, Jin Hyeon Kim, Paul Hyunbin Cho, Jaeeun Lee, Jihye Park, Minkyu Park, Sangpil Kim, Hyunhee Park, and Seungryong Kim. Text-aware image restoration with diffusion models. In *ICLR*, 2026.

[12] Haodong He, Xin Zhan, Yancheng Bai, Rui Lan, Lei Sun, and Xiangxiang Chu. Texts-diff: Texts-aware diffusion model for real-world text image super-resolution. In *ICASSP*, 2026.

[13] Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *ICCV*, 2021.

[14] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *ICCVW*, 2021.

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[16] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *IJCV*, 2024.

[17] 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 *ECCV*, 2024.

[18] Xiaoyang Liu, Yuquan Wang, Zheng Chen, Jiezhang Cao, He Zhang, Yulun Zhang, and Xiaokang Yang. One-step diffusion model for image motion-deblurring. *arXiv preprint arXiv:2503.06537*, 2025.

[19] Xiaoyang Liu, Zhengyan Zhou, Zihang Xu, Jiezhang Cao, Zheng Chen, and Yulun Zhang. Fidediff: Efficient diffusion model for high-fidelity image motion deblurring. In *ICLR*, 2026.

[20] Rongyuan Wu, Tao Yang, Lingchen Sun, Zhengqiang Zhang, Shuai Li, and Lei Zhang. Seesr: Towards semantics-aware real-world image super-resolution. In *CVPR*, 2024.

[21] 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 *CVPR*, 2024.

[22] Zheng Chen, Bowen Chai, Rongjun Gao, Mingtao Nie, Xi Li, Bingnan Duan, Jianping Fang, Xiaohong Liu, Linghe Kong, and Yulun Zhang. Dvface: Spatio-temporal dual-prior diffusion for video face restoration. *arXiv preprint arXiv:2604.14560*, 2026.

[23] Zheng Chen, Ruofan Yang, Jin Han, Dehua Song, Zichen Zou, Chunming He, Yong Guo, and Yulun Zhang. Diffst: Spatiotemporal-aware diffusion for real-world space-time video super-resolution. *arXiv preprint arXiv:2605.13182*, 2026.

[24] 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 *CVPR*, 2024.

[25] Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *CVPR*, 2025.

[26] Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *ACM TOG*, 2025.

[27] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024.

[28] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[29] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[30] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[31] 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 *ICCV*, 2025.

[32] Jianqi Ma, Shi Guo, and Lei Zhang. Text prior guided scene text image super-resolution. *IEEE TIP*, 2023.

[33] Minyi Zhao, Shijie Xuyang, Jihong Guan, and Shuigeng Zhou. Stirer: A unified model for low-resolution scene text image recovery and recognition. In *ACM MM*, 2023.

[34] Hang Guo, Tao Dai, Guanghao Meng, and Shu-Tao Xia. Towards robust scene text image super-resolution via explicit location enhancement. In *IJCAI*, 2023.

[35] Xiangyuan Zhu, Kehua Guo, Hui Fang, Rui Ding, Zheng Wu, and Gerald Schaefer. Gradient-based graph attention for scene text image super-resolution. In *AAAI*, 2023.

[36] Shipeng Zhu, Zuoyan Zhao, Pengfei Fang, and Hui Xue. Improving scene text image super-resolution via dual prior modulation network. In *AAAI*, 2023.

[37] Zuoyan Zhao, Hui Xue, Pengfei Fang, and Shipeng Zhu. Pean: A diffusion-based prior-enhanced attention network for scene text image super-resolution. In *ACM MM*, 2024.

[38] Baole Wei, Yuxuan Zhou, Liangcai Gao, and Zhi Tang. Glyphsr: A simple glyph-aware framework for scene text image super-resolution. In *AAAI*, 2025.

[39] Xiaoming Li, Wangmeng Zuo, and Chen Change Loy. Enhanced generative structure prior for chinese text image super-resolution. *IEEE TPAMI*, 2025.

[40] Shrey Singh, Prateek Keserwani, Masakazu Iwamura, and Partha Pratim Roy. Dcdm: Diffusion-conditioned-diffusion model for scene text image super-resolution. In *ECCV*, 2024.

[41] Jianze Li, Jiezhang Cao, Zichen Zou, Xiongfei Su, Xin Yuan, Yulun Zhang, Yong Guo, and Xiaokang Yang. Unleashing the power of one-step diffusion based image super-resolution via a large-scale diffusion discriminator. In *NeurIPS*, 2025.

[42] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[43] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: similarity control and knowledge transfer. *JMLR*, 2015.

[44] Wonkyung Lee, Junghyup Lee, Dohyung Kim, and Bumsub Ham. Learning with privileged information for efficient image super-resolution. In *ECCV*, 2020.

[45] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. In *ICCV*, 2023.

[46] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

[47] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

[48] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017.

[49] Qian Ning, Weisheng Dong, Xin Li, Jinjian Wu, and Guangming Shi. Uncertainty-driven loss for single image super-resolution. In *NeurIPS*, 2021.

[50] Zhenxuan Fang, Fangfang Wu, Weisheng Dong, Xin Li, Jinjian Wu, and Guangming Shi. Self-supervised non-uniform kernel estimation with flow-based motion prior for blind image deblurring. In *CVPR*, 2023.

[51] Jianqi Ma, Zhetong Liang, Wangmeng Xiang, Xi Yang, and Lei Zhang. A benchmark for chinese-english scene text image super-resolution. In *ICCV*, 2023.

[52] Haiyang Yu, Jingye Chen, Bin Li, Jianqi Ma, Mengnan Guan, Xixi Xu, Xiaocong Wang, Shaobo Qu, and Xiangyang Xue. Benchmarking chinese text recognition: Datasets, baselines, and an empirical study. *arXiv preprint arXiv:2112.15093*, 2021.

[53] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *ICCV*, 2021.

[54] 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 *CVPRW*, 2022.

[55] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023.

[56] 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 *ICLR*, 2022.

[57] Jingkai Wang, Jue Gong, Lin Zhang, Zheng Chen, Xing Liu, Hong Gu, Yutong Liu, Yulun Zhang, and Xiaokang Yang. Osdface: One-step diffusion model for face restoration. In *CVPR*, 2025.

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[59] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

[60] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025.

A Details of BTL Dataset Construction

Motivation and source data. Existing text-image datasets cover different aspects of Text-SR, but no single resource fully matches our training setting of high-quality Chinese-English text-line super-resolution. TextZoom [5] provides real paired LR-HR scene text images and has been widely used as a benchmark for scene Text-SR, but it mainly focuses on English text and often suffers from varying image quality. RealCE [51] further introduces a Chinese-English scene Text-SR benchmark with an emphasis on structurally complex Chinese characters, but its scale is relatively limited.

For constructing BTL, we use two annotated text-image sources as real-image candidate pools for HQ text-line crops. The first is CTR [52], a large-scale Chinese text recognition benchmark built from multiple scene-text datasets. Although CTR is designed for recognition rather than super-resolution, it provides a large number of cropped text-line images with transcripts and contains a high proportion of Chinese samples, making it suitable for selecting Chinese real-text candidates. The second is SA-Text [11], a large-scale text-aware image restoration dataset built from high-quality scene images with detailed text annotations. Since SA-Text provides abundant English text instances and high-quality visual content, we use it as the main source for English real-text candidates. In addition, we generate synthetic HQ text-line images following the rendering strategy of MARCONet [4]. The final BTL dataset combines curated real HQ crops from CTR and SA-Text with synthetic HQ text-line images, balancing real-world appearance, language coverage, and controllable text-line diversity.

Language statistics of source pools. We analyze the transcript distribution of the two real-image source pools used for BTL construction, i.e., CTR [52] and SA-Text [11]. We categorize each candidate according to its transcript. *Chinese* denotes samples containing at least one Chinese character; *English* denotes samples containing English letters but no Chinese characters; *Digit* denotes samples containing only digits after removing punctuation; and the remaining samples are grouped as *Others*. As shown in Tab. 4, CTR contains a substantially larger proportion of Chinese text, while SA-Text provides more English candidates. This supports our source allocation strategy: Chinese candidates are selected from CTR, whereas English candidates are selected from SA-Text.

Real HQ text-line crop selection. All candidate crops from CTR [52] and SA-Text [11] are filtered using the following protocol. First, each crop is resized to a fixed height of 128 using bicubic interpolation while preserving its aspect ratio, so that all candidates are assessed under a consistent resolution. We then retain samples whose aspect ratio falls between 2 and 8 and whose transcript length is no more than 24 characters. These constraints remove unsuitable text-line geometries, such as extremely short, overly long, or densely annotated instances, and match our target setting of high-quality bilingual text-line SR. Finally, we rank the retained candidates using no-reference IQA metrics to select visually reliable crops. Within each language/source group, MUSIQ, MANIQA, and CLIP-IQA scores are converted into percentile ranks in $[0, 1]$, denoted as $\mathcal{R}_{\mathrm{MUSIQ}}$, $\mathcal{R}_{\mathrm{MANIQA}}$, and $\mathcal{R}_{\mathrm{CLIP-IQA}}$, respectively. The final quality score is computed as

$$\mathcal{Q} = 0.50\mathcal{R}_{\mathrm{MUSIQ}} + 0.35\mathcal{R}_{\mathrm{MANIQA}} + 0.15\mathcal{R}_{\mathrm{CLIP-IQA}}.$$

We sort candidates by $\mathcal{Q}$ within each group and allocate the selection quota according to the group proportion in the retained candidate pool, preserving the original group distribution in the curated subset. The resulting curated real HQ subset contains 50K images, including 31,567 Chinese samples and 1,084 digit-only samples from CTR, and 16,634 English samples and 715 digit-only samples from SA-Text.

Synthetic HQ text-line images. To increase scale and content diversity, we additionally generate 50K synthetic HQ text-line images following the rendering strategy of MARCONet [4]. Rendered samples provide controllable high-quality text-line images with diverse transcripts, fonts, and layouts, while the curated real subset contributes real image statistics and background-text interactions. Combining these two sources allows BTL to balance controllability and real-world appearance.

LQ synthesis and dataset split. For each HQ text-line image, we synthesize the corresponding LQ input using degradation pipelines based on BSRGAN [13] and Real-ESRGAN [14]. The final BTL dataset contains 100K HQ images, consisting of 50K curated real crops and 50K rendered text-line images. We preserve the real/synthetic ratio and split BTL into 80K training samples and 20K testing samples, denoted as BTL-train and BTL-test, respectively. Tab. 5 summarizes the final dataset composition.

Source	Total	Chinese	English	Digit	Others
CTR	636,455	519,974	82,484	31,188	2,809
SA-Text	271,126	1,743	236,258	29,336	3,789

Table 4: Language statistics of the real-image source pools used for BTL construction.

Subset	Curated Real HQ	Synthetic HQ	Total
BTL-train	40,000	40,000	80,000
BTL-test	10,000	10,000	20,000
Total	50,000	50,000	100,000

Table 5: Final composition of BTL. For each HQ image, the corresponding LQ input is synthesized using the degradation pipelines described in Sec. A.

Effect of data composition. We further examine the effect of different HQ data sources by training PRISM with three data configurations and evaluating them on RealCE-val. The first configuration, denoted as Synth-train, uses synthetic text lines generated following the rendering strategy of MARCONet. The second configuration, denoted as CTR-train, uses real text crops from the training split of CTR. The third configuration is the proposed BTL-train, which combines curated real crops and rendered text lines. All three PRISM variants are trained with the same training configuration and evaluated on the same RealCE-val set. As shown in Fig. 9, PRISM trained only on Synth-train tends to produce sharp but less naturally integrated strokes in some real-world cases, where the restored text may appear separated from the background. PRISM trained only on CTR-train produces more conservative results, but the restored text is often less clear. In comparison, PRISM trained on BTL-train provides a better balance between text sharpness and real-world appearance. This qualitative analysis suggests that combining curated real crops and rendered text lines is beneficial for real-world Text-SR.

B RealCE-val Evaluation Protocol

We construct RealCE-val from the original text-line pairs in RealCE [51]. The 52 mm crops serve as HQ references, while the corresponding 26 mm and 13 mm crops are used as $\times 2$ and $\times 4$ LQ inputs, respectively. Before filtering, we manually verify the transcripts against the HQ images and correct erroneous annotations.

We then exclude samples based on the following criteria: (i) missing or unreadable crops; (ii) an aspect ratio below 0.5 for the HQ reference; (iii) failed ECC-based Euclidean registration between the HQ and LQ crops; (iv) an ECC-estimated translation exceeding 3.5 pixels; (v) a normalized three-channel color-histogram correlation below 0.5; or (vi) an absolute difference in mean image intensity exceeding 10. The same criteria and thresholds apply to both $\times 2$ and $\times 4$ pairs.

All filtering criteria and thresholds are fixed before model evaluation. The resulting RealCE-val contains 1,037 valid pairs, with identical samples, corrected transcripts, preprocessing, and evaluation metrics used for all compared methods.

C Inference Speed Analysis

Table 6 presents a detailed inference speed comparison on RealCE-val among different methods under the $\times 4$ setting, including the number of sampling steps and per-image inference time. For a fair comparison, we select images from RealCE-val whose heights range from 50 to 100 pixels and whose widths range from 300 to 500 pixels and then resize them to the target resolution. For non-diffusion-based methods, we measure runtime with a fixed LR input size of 32×128 and a fixed HR output size of 128×512. For diffusion-based methods, both the input and output are fixed to 128×512. All methods are evaluated with batch size 1 on a single RTX 4090, and the measured runtime includes the full inference pipeline except file I/O. As shown in Tab. 6, PRISM achieves substantially faster inference than existing diffusion-based Text-SR methods. Compared with DiffTSR [9] which uses 200 sampling steps, and TeReDiff [11] which uses 50 steps, PRISM

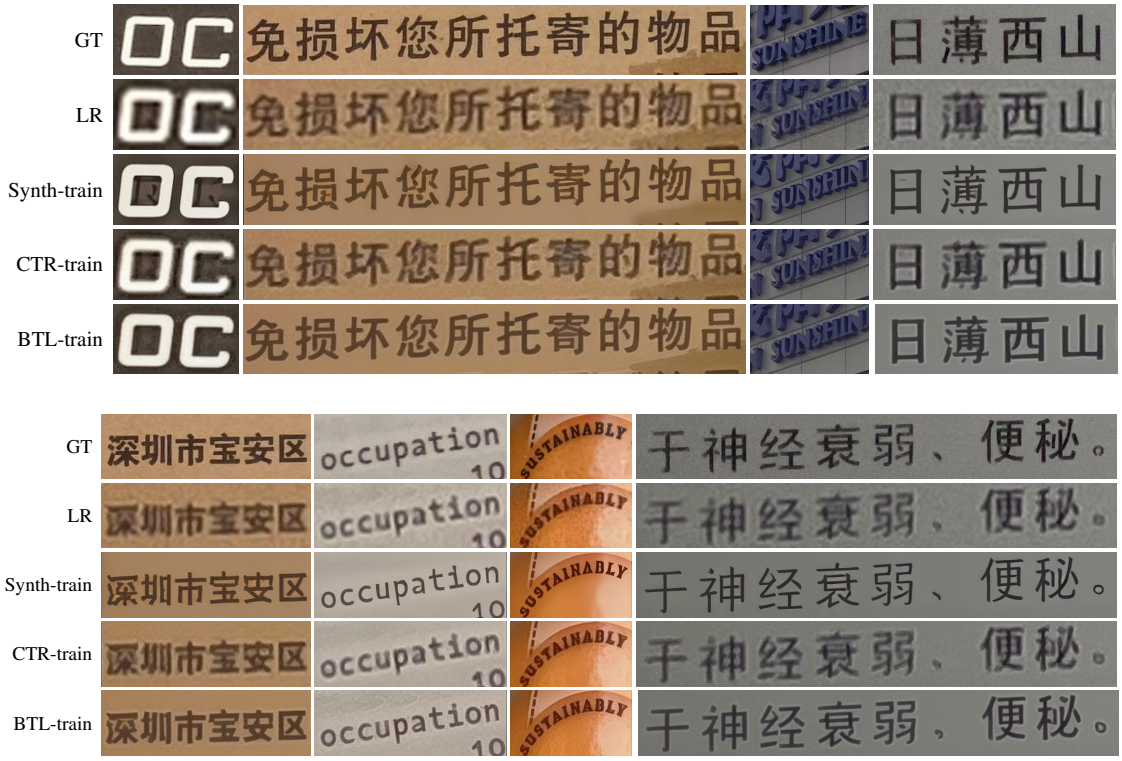

Figure 9: Effect of training data composition on RealCE-val for $\times 4$ super-resolution. We compare PRISM trained with three data configurations: Synth-train, CTR-train, and BTL-train.

	TSRN	TBSRN	TATT	MARCONet	DiffTSR-s200	TeReDiff-s50	**Ours**
Inference Time / ms	6.14	10.48	25.94	320.11	10702.98	5273.42	**84.74**

Table 6: Inference speed comparison ($\times 4$) on RealCE-val among different methods. Diffusion-based methods use different numbers of inference steps: DiffTSR [9] uses 200 steps, TeReDiff [11] uses 50 steps, and ours uses only one step.

requires only one step and takes around 80 milliseconds per image. Notably, its runtime is comparable to the non-diffusion-based Text-SR methods. These results show that PRISM retains restoration capability while achieving practical one-step inference efficiency.

D More Visualizations

We provide additional visual comparisons on the synthetic BTL-test set and the real-world RealCE-val set in Figs. 10 and 11. These examples cover diverse text-line types, including Chinese, English, and digit-only samples, as well as both short and long text lines. As shown in the figures, TSRN [5], TBSRN [6], and TATT [7] often produce relatively smooth results, especially for structurally complex Chinese characters, where fine strokes and character components are difficult to recover. MARCONet [4] generally improves visual sharpness, but it may introduce distorted glyph shapes, unnatural foreground-background separation, or local stroke artifacts such as broken and merged strokes. DiffTSR [9] can restore plausible text structures in some cases, but under severe degradation it may miss fine strokes or produce incorrect characters. TeReDiff [11] produces sharp outputs, yet it sometimes introduces unrealistic textures, color artifacts, or extra strokes, particularly on small or heavily degraded text images. In contrast, PRISM restores clearer and more coherent text structures across both synthetic and real-world examples. The characters recovered by PRISM show fewer broken or merged strokes, and the text regions are better integrated with the surrounding background. These visual results further indicate that the proposed recoverable prior and uncertainty-aware structure modeling help improve both perceptual clarity and text readability under challenging degradation.

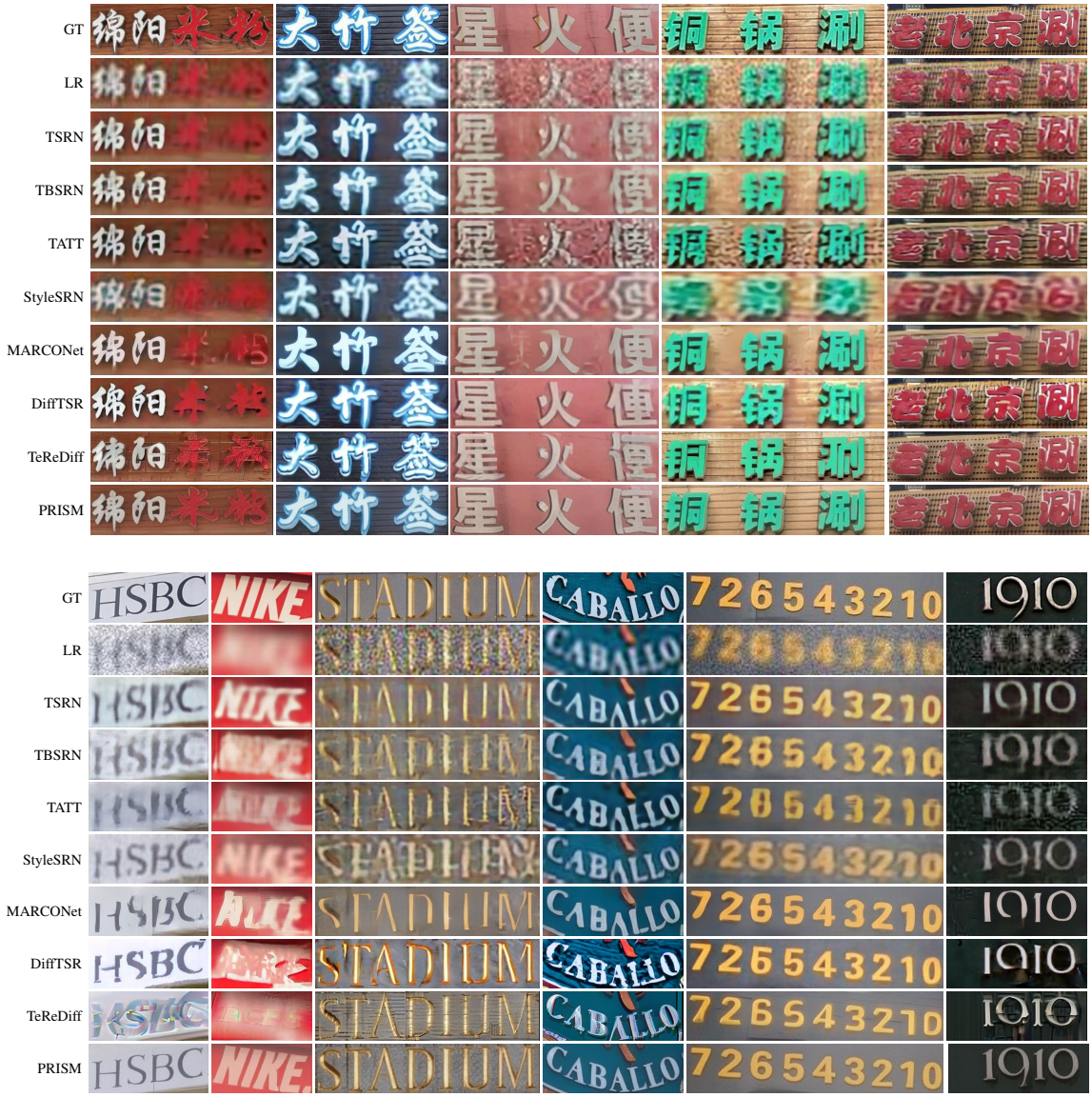

Figure 10: More visualizations on the synthetic dataset BTL-test for $\times 4$ super-resolution.

E Architecture Details of PRISM

Prior Encoder in FMPR. The privileged prior encoder $\mathcal{E}_{\mathrm{p}}$ and the LQ-only prior encoder $\mathcal{E}_{\mathrm{lq}}$ share the same architecture except for the input channel number. The input is a latent tensor $z \in \mathbb{R}^{C_{\mathrm{in}} \times \frac{H}{8} \times \frac{W}{8}}$, where $C_{\mathrm{in}} = 8$ for $\mathcal{E}_{\mathrm{p}}$ because it takes the channel-wise concatenation of z_l and z_h, and $C_{\mathrm{in}} = 4$ for $\mathcal{E}_{\mathrm{lq}}$ because it only takes the LQ latent z_l. The encoder first maps the input to a 256-channel feature space with a 3×3 convolution and LeakyReLU, followed by four residual blocks at the same spatial resolution. The feature is then projected to 1024 channels by three 3×3 convolution layers and adaptively pooled to a fixed spatial size of 4×16. This yields $N = 64$ spatial tokens. After reshaping the feature into a sequence of shape $(B, 64, 1024)$, a two-layer MLP mixer performs token-wise and channel-wise mixing with LayerNorm. A final linear projection produces the prior embedding in $\mathbb{R}^{B\times64\times1024}$.

Flow-Matching Velocity Network in FMPR. The velocity network $\mathcal{V}_{\mathrm{FM}}$ is a lightweight token-wise MLP operating in the prior embedding space. Given the current prior embedding $c^k \in \mathbb{R}^{B\times N\times D}$ and the normalized integration step k/K, the scalar timestep is broadcast to $(B, N, 1)$ and concatenated with c^k along the feature dimension. A linear layer first maps the resulting $(B, N, D + 1)$ representa-



Figure 11: More visualizations on the real-world dataset RealCE-val for $\times 4$ super-resolution.

tion back to dimension $D = 1024$. The representation is then refined by four residual MLP blocks, each consisting of a linear layer and LeakyReLU activation. The network outputs a velocity tensor in $\mathbb{R}^{B\times N\times D}$, which is used in the $K = 16$ step Euler integration in Eq. (5).

Uncertainty-Aware Spatial Cue Extractor in SURE. The spatial cue extractor $\mathcal{F}_\eta$ takes the degraded image $x_l \in \mathbb{R}^{B\times 3\times H\times W}$ as input and produces a projected structural cue $p_s \in \mathbb{R}^{B\times 320\times \frac{H}{8}\times \frac{W}{8}}$ for the structural residual encoder. It also predicts an auxiliary boundary map $\hat{m} \in \mathbb{R}^{B\times 1\times H\times W}$ for structure supervision.

The extractor consists of a convolutional stem, five downsampling blocks, and a lightweight Feature Pyramid Network (FPN) for top-down fusion. The stem maps the input image to 32 channels with a 3×3 convolution, GroupNorm, SiLU, and a residual block. The following downsampling blocks gradually reduce the spatial resolution and produce multi-scale features with channel dimensions $(32, 64, 128, 256, 512)$. A lightweight FPN fuses the last three scales in a top-down manner and produces an $\frac{H}{8} \times \frac{W}{8}$ feature map $p_{\text{raw}} \in \mathbb{R}^{B\times 128\times \frac{H}{8}\times \frac{W}{8}}$. The uncertainty-aware latent head operates on p_{raw}. Two parallel convolutional heads predict the mean μ and log-variance $\log \sigma^2$ of a latent structural cue distribution. During training, the stochastic structural cue is sampled by the reparameterization in Eq. (7). The sampled feature is projected to p_s through a learnable projection layer and sent to the structural residual encoder. In parallel, an edge head decodes the sampled feature into the auxiliary boundary map $\hat{m}$. During inference, we use a noise-attenuated stochastic cue derived from the predicted distribution.

Structural Residual Encoder in SURE. The structural residual encoder $\mathcal{C}_\eta$ follows a ControlNet-style residual conditioning design and is initialized from the encoder part of the diffusion backbone.

Methods	Updates	Time Cond.	Params (M)	FLOPs (G)	LPIPS ↓	ACC ↑	NED ↑
Direct regression	1	–	66.79	69.44	0.2120 ± 0.0021	$63.97\% \pm 0.11\%$	0.8435 ± 0.0014
Recurrent refinement	16	0	30.05	74.81	0.2152 ± 0.0014	$63.90\% \pm 0.19\%$	0.8423 ± 0.0004
FMPR	16	k/K	30.05	74.81	$\mathbf{0.2059 \pm 0.0007}$	$\mathbf{64.58\% \pm 0.06\%}$	$\mathbf{0.8459 \pm 0.0005}$

Table 7: Comparison of prior recovery paradigms on RealCE-val ($\times 4$). Results are reported as mean $\pm$ standard deviation over three independent seeds.

Methods	PSNR ↑	LPIPS ↓	ACC ↑	NED ↑
w/o uncertainty	19.8763 ± 0.0012	0.2055 ± 0.0006	$64.80\% \pm 0.06\%$	0.8478 ± 0.0003
Full model	$\mathbf{19.8920 \pm 0.0009}$	$\mathbf{0.2041 \pm 0.0005}$	$\mathbf{65.16\% \pm 0.06\%}$	$\mathbf{0.8515 \pm 0.0009}$

Table 8: Multi-seed stability of SURE on RealCE-val ($\times 4$). Results are reported as mean $\pm$ standard deviation over three independent seeds.

It takes the degraded latent z_l, the recovered prior $\hat{c}$, and the projected structural cue p_s as inputs, and predicts multi-level residual controls $\mathcal{R} = \{r_i\}_{i=1}^{M}$, where M equals 9 according to the UNet structure. These residuals are injected into the skip-connection features of the frozen UNet $\mathcal{U}_{\bar{\theta}}$. Since both the FMPR pathway and the restoration backbone are frozen in this stage, $\mathcal{C}_\eta$ focuses on residual spatial refinement rather than re-estimating the text-aware prior.

F Additional Ablation Studies

FMPR: Compute-Matched Recurrent Refinement and Stability. We further investigate whether the improvement of FMPR stems from time-conditioned prior rectification or merely repeated refinement. We introduce a 16-step recurrent refinement variant with the same velocity network architecture, parameters, FLOPs and training settings as FMPR, except that the normalized timestep k/K is replaced with a constant zero. Direct regression is retained as a single-pass reference.

As shown in Tab. 7, FMPR outperforms both direct regression and the compute-matched recurrent refinement variant across three independent seeds. The consistent gains over recurrent refinement indicate that time-conditioned transport contributes beyond repeated updates alone.

SURE: Multi-Seed Stability. We further evaluate the stability of SURE by comparing the deterministic structure guidance (w/o uncertainty) with the full model over three independent seeds.

As shown in Tab. 8, the full model consistently outperforms deterministic guidance across the reported metrics, supporting the stability of its gains.

G Broader Impacts and Limitations

Broader impacts. This work aims to improve the readability and visual quality of degraded text images. PRISM may benefit applications such as document enhancement, scene text recognition, assistive reading, and OCR preprocessing. Since text super-resolution may reconstruct plausible content from ambiguous inputs, restored results should be used with caution in sensitive scenarios. For legal, medical, financial, or privacy-related use cases, they should be regarded as auxiliary references rather than authoritative evidence.

Limitations. Our study focuses on Chinese-English text-line super-resolution with moderate to long aspect ratios. This setting covers many practical text-image cases, but does not fully include full-scene text restoration, dense multi-line documents, or highly irregular text layouts. Extending PRISM to broader text-aware restoration scenarios is a promising future direction. In addition, although BTL combines curated real crops and rendered text lines, its language coverage is still mainly limited to Chinese, English, and digit-based text. Future work may further expand the dataset to more languages, scripts, fonts, and real-world capture conditions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All the claims in the abstract and the introduction precisely reflect the paper's contribution and are supported by the experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the paper's limitations in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided experiment details in the implementation details section and dataset details in the appendix for result reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code will be released at `https://github.com/faithxuz/PRISM`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have included the training and test details in the corresponding section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report mean and standard deviation over three independent training seeds for FMPR and SURE in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report it in the implementation details section and the inference efficiency section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research fully complies with the NeurIPS Code of Ethics in all respects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss it in the appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the works we mentioned or used and fully respect their license.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The construction of BTL and the RealCE-val evaluation protocol are described in the appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLM is used only for writing and editing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.