# A Bi-Objective Framework for Scene Text Super-Resolution: Balancing Legibility and Fidelity

**Anonymous authors**
**Paper under double-blind review**

## Abstract

Image super-resolution (SR) can improve many vision systems—from surveillance and autonomy to document analysis and retail analytics—because recovering high-frequency details, especially scene text, enables reliable downstream perception. Scene text, i.e., text embedded in natural images such as signs, product labels, and storefronts, often carries task-critical information; when characters are blurred or hallucinated, optical character recognition (OCR) and subsequent decisions can fail even if the rest of the image appears sharp. Yet previous SR research has often been tuned to distortion metrics such as PSNR/SSIM or perceptual and image-quality metrics such as LPIPS, MANIQA, CLIP-IQA, and MUSIQ, none of which directly measure character-level correctness. Furthermore, studies that do address text SR often rely on constrained text-centric benchmarks, leaving text restoration in complex natural scenes underexplored. As a result, scene text is effectively treated as generic texture. For SR to be effective in practical deployments, it is therefore essential to explicitly optimize for both text legibility and perceptual quality. We present GLYPH-SR, a vision–language-guided diffusion framework that aims to achieve both objectives jointly. GLYPH-SR uses a Text-SR Fusion ControlNet (TS-ControlNet) guided by OCR-derived text and layout cues, together with a ping–pong scheduler that alternates between text-centric and scene-centric guidance. To enable targeted text restoration, we train TS-ControlNet on a synthetic corpus while keeping the main SR branch frozen. Across SVT, SCUT-CTW1500, and CUTE80 at ×4 and ×8, GLYPH-SR improves OCR $F_1$ by up to 15.18 percentage points over diffusion/GAN baselines (SVT ×8, OpenOCR) while maintaining competitive MANIQA, CLIP-IQA, and MUSIQ. GLYPH-SR is designed to better balance the two objectives—text readability and visual realism—producing outputs that are both perceptually plausible and more text-legible. We provide code, pretrained models, the synthetic corpus with generation scripts, and an evaluation suite to support reproducibility.

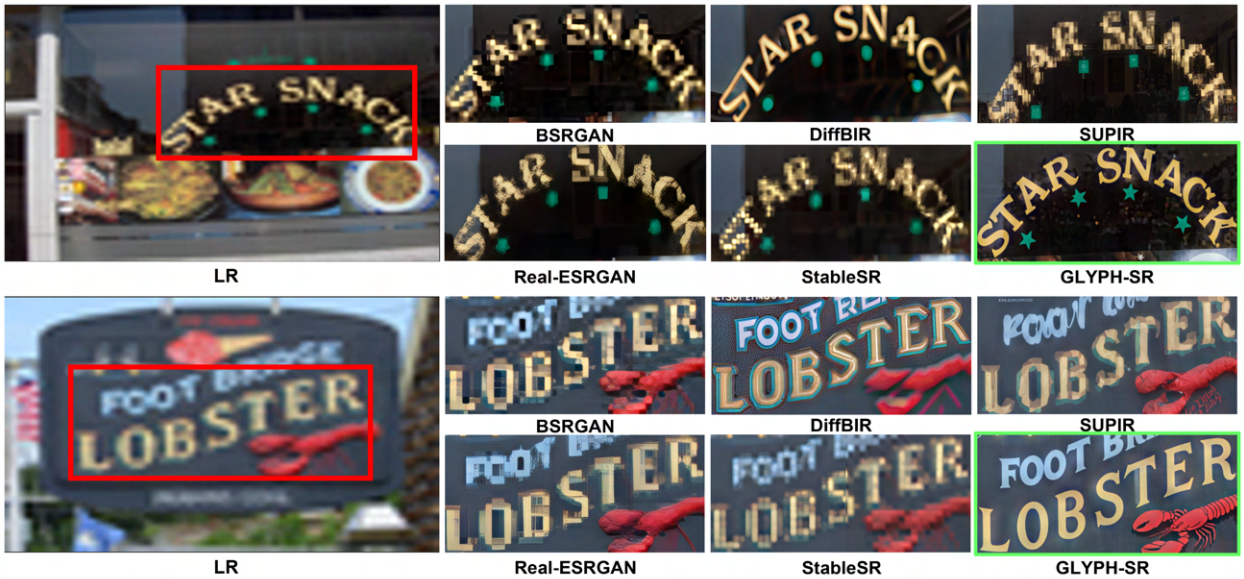
## 1 Introduction

Image super-resolution (SR),[1] which reconstructs high-resolution (HR) images from low-resolution (LR) inputs, can support applications such as autonomous driving, document analysis, and retail analytics, where fine visual details affect downstream perception. While conventional SR aims to improve perceptual quality, we argue that for many real-world scenarios, ensuring the legibility of scene text (e.g., on signs, license plates) can be at least as important. Accurately restoring characters is crucial, as failures in legibility can compromise downstream tasks like optical character recognition (OCR), regardless of the overall image sharpness.

### 1.1 Scene Text Fidelity: An Overlooked Challenge in Image SR

However, achieving this level of text fidelity remains an overlooked challenge in most conventional SR frameworks. Two systemic biases explain why text often degrades in existing SR models (e.g., StableSR Wang et al. (2024), DiffBIR Lin et al. (2024), InvSR Yue et al. (2025)) despite strong perceptual scores:

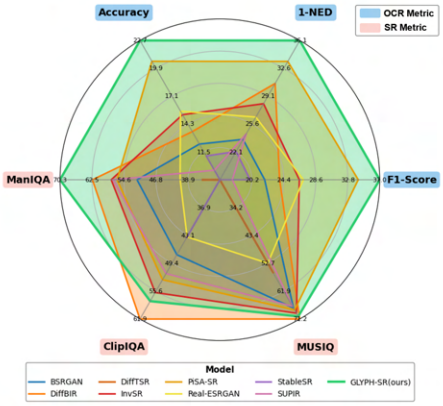
[1]Throughout this paper, we use *image SR* and *SR* interchangeably when unambiguous.




(a) Compare diffusion result with SR models

| Method | Image Fidelity | Text Restoration | Text Fidelity |
|---|---|---|---|
| SUPIR | ✓ | ✓ | △ |
| BSRGAN | ✓ | ✗ | ✗ |
| Real-ESRGAN | ✓ | ✓ | ✗ |
| StableSR | ✓ | ✓ | ✗ |
| DiffBIR | ✓ | ✓ | △ |
| STISR | ✗ | ✓ | ✗ |
| TATT | ✗ | ✓ | ✗ |
| SeeSR | ✓ | ✓ | △ |
| InvSR | ✓ | ✓ | △ |
| DiffTSR | △ | ✓ | ✓ |
| PiSA-SR | ✓ | ✓ | △ |
| **Ours** | ✓ | ✓ | ✓ |

(b) Comparison of recent SR methods across three qualitative criteria: image fidelity, text restoration, and text fidelity. ✓ = Supported, ✗ = Not supported, △ = Not enough.




(c) Evaluate of SR models on SCUT-CTW1500(x4) with respect to SR and OCR metrics.

Figure 1: Qualitative and quantitative comparisons between GLYPH-SR and competing SR methods, demonstrating superior text fidelity and OCR $F_1$ score.

(a) **Metric Bias.** Standard full-reference distortion metrics (PSNR/SSIM) and learned/no-reference perceptual metrics (LPIPS, MANIQA, CLIP-IQA, MUSIQ) aggregate quality globally and are dominated by area; because small text regions often occupy only a small fraction of the image, their character-level corruptions are weakly penalized.

(b) **Objective Bias.** Common training losses prioritize appearance similarity and treat characters as generic high-frequency texture rather than discrete semantic units required by OCR.

In practice these biases surface as two failure modes (Fig. 1 (a)): *(i) Hallucination*—methods optimized for perceptual realism may produce sharp but incorrect characters, harming OCR; *(ii) Conservative restoration*—others preserve the blurry input to avoid artifacts, yielding limited SR gains and limited perceptual improvement.

As a result, few approaches simultaneously enhance visual realism and ensure text legibility—an essential requirement for OCR-dependent applications.

### 1.2 Contributions

We address scene text SR as a *bi-objective* problem—optimizing both **visual quality** and **text legibility**—and present **GLYPH-SR**, a vision–language-guided diffusion framework for jointly improving both objectives. Our key contributions are as follows:

- **Bi-Objective Formulation & Dual-Axis Evaluation.** We explicitly cast SR in text-rich scenes as the joint optimization of *image quality* and *readability*, and adopt a *dual-axis* protocol that reports perceptual SR metrics (MANIQA, CLIP-IQA, MUSIQ) *together with* OCR-aware measures (word/character accuracy, edit distance, $F_1$), ensuring that small text regions are not underweighted.
- **Text-SR Fusion ControlNet with Time-Balanced Guidance.** We introduce a dual-branch **TS-ControlNet** that fuses *token-level OCR strings with verbalized locations* $\mathcal{S}_{\text{TXT}}$ and a *scene caption* $\mathcal{S}_{\text{IMG}}$. The SR branch is frozen while the text branch is fine-tuned; residual mixing injects complementary cues into the LDM without disrupting its generative prior. A lightweight **ping–pong** scheduler $\lambda_t$ alternates text-centric and image-centric conditioning along the denoising trajectory, and coherently modulates both *embedding fusion* and *residual injection*.
- **Factorized Synthetic Corpus & Comprehensive Validation.** We build a four-partition synthetic corpus that independently perturbs glyph quality and global image quality, enabling targeted text restoration while keeping the SR branch frozen. Across SVT, SCUT-CTW1500, and CUTE80 at $\times 4/\times 8$, **GLYPH-SR** improves OCR $F_1$ by up to **15.18 percentage points** over strong diffusion/GAN baselines while maintaining competitive MANIQA, CLIP-IQA, and MUSIQ. We release code, pretrained models, data-generation scripts, and an evaluation suite to support reproducibility.

## 2 Related Work

**SR via Deep Learning.** Early CNN-based methods such as SRCNN Dong et al. (2015), EDSR Lim et al. (2017), and RCAN Zhang et al. (2018b), followed by Transformer-based models such as SwinIR Liang et al. (2021), have substantially advanced distortion-oriented SR. However, these methods primarily optimize pixel-level fidelity, typically measured by PSNR and SSIM, rather than text legibility or character-level correctness in small text-containing regions. Adversarially trained SR methods improve perceptual realism on in-the-wild images, with representative examples including BSRGAN Zhang et al. (2021) and Real-ESRGAN Wang et al. (2021).

**General-Purpose SR Models.** Diffusion-based SR has recently shown strong perceptual realism by leveraging powerful generative priors. Representative approaches such as DiffBIR Lin et al. (2024), ConsisSR Gu et al. (2024), and StableSR Wang et al. (2024) combine LR conditioning with diffusion priors, and subsequent methods incorporate richer priors or auxiliary conditions. Some methods further exploit text-based prompts: SeeSR Wu et al. (2024) uses semantic prompts, while PromptSR Chen et al. (2023) injects text prompts to improve restoration performance. SUPIR Yu et al. (2024b) also leverages text prompts, combined with large-scale pretrained backbones and restoration-guided sampling. Other approaches include InvSR Yue et al. (2025), which enables flexible guidance and sampling, and PISA-SR Sun et al. (2025), which further improves controllability. Nevertheless, explicit character-level integrity is rarely a primary optimization target in general-purpose diffusion SR. This leaves a gap between perceptual realism and text-specific restoration quality.

**Text-Focused SR.** Text-centric SR aims to improve readability using text-aware priors or recognition-aware objectives. Representative methods include TATT Ma et al. (2022), STISR Noguchi et al. (2024), MARCONet Li et al. (2023; 2025), and Stroke-Aware SR Chen et al. (2022). Most of these methods focus on cropped word- or line-level images, where text is the primary visual content and its location is implicitly given. Full-scene text SR differs from cropped text SR because it operates on larger images, where text occupies only a small portion of the scene and must be restored without degrading non-text regions. Therefore, this

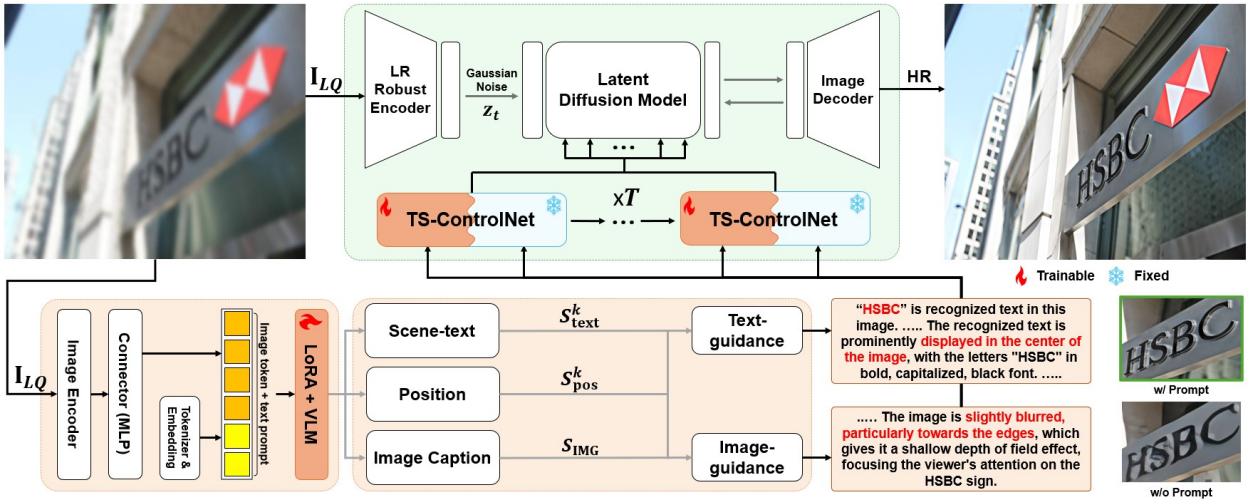



Figure 2: Overview of the proposed GLYPH-SR architecture.

task demands a different trade-off to preserve global perceptual quality. More recent work, such as Min et al. (2025), considers scene text restoration, but handling severe low-resolution degradation and the joint optimization of readability and realism remain underexplored.

Overall, scene text SR in full natural images remains underexplored, particularly when both global perceptual fidelity and precise text restoration must be preserved under severe LR conditions.

## 3 Our Approach: GLYPH-SR

### 3.1 Model Architecture

**Overview.** Fig. 2 depicts the proposed **GLYPH-SR** pipeline. Given an LR image $\mathbf{I}_{\text{LR}} \in \mathbb{R}^{H\times W\times C}$, an LR-robust conditioner of a pretrained latent diffusion model (LDM) Rombach et al. (2022) extracts multi-scale LR features $f_{\text{LR}}$ for restoration conditioning. Our **Text–SR Fusion ControlNet (TS-ControlNet)** injects complementary text-aware restoration cues while preserving the generative prior of the LDM. The reverse diffusion process is then performed in latent space to obtain a high-resolution reconstruction. While this LR-conditioned diffusion pipeline provides strong global restoration, holistic guidance alone is often insufficient for small text regions. Such regions can be treated as generic high-frequency textures rather than semantically meaningful glyphs, leading to imperfect character restoration.

**Condition Decomposition.** To address this limitation, we explicitly decompose the guidance into **image-oriented** and **text-oriented** signals.

- **Image-Oriented Guidance.** A scene-level caption $\mathcal{S}_{\text{IMG}}$ summarizes global scene attributes, such as content, layout, and visual context, and is used to encourage holistic perceptual consistency.
- **Text-Oriented Guidance.** A dedicated OCR module detects $K$ text instances and produces OCR-derived text–position pairs $\mathcal{S}_{\text{TXT}} = \{(s^k_{\text{text}}, s^k_{\text{pos}})\}_{k=1}^{K}$ [2]. Each pair is converted into a structured natural-language prompt, e.g., "`HSBC is displayed at the center of the image`," and passed to the text branch.

As shown in Fig. 3(a), using text-oriented guidance alone can improve text fidelity but may degrade non-text regions. This motivates our guidance-fusion strategy and the ping–pong scheduler, which balance text-centric and scene-centric conditioning during the denoising process.

[2]At test time, all text strings and positions are obtained from the upstream OCR/VLM module applied to the LR input. Ground-truth transcriptions or annotations from the evaluation datasets are never used as guidance.

Figure 3: Text-centric fine-tuning framework: (a) trade-off between scene text fidelity and overall image quality under different guidance strategies; (b) four synthetic training subsets with matched prompts; (c) TS-ControlNet architecture.

**Text–SR Fusion ControlNet.** To balance the two objectives—image quality and text legibility—we introduce the *Text-SR Fusion ControlNet* (TS-ControlNet), which combines glyph-level semantic priors with global SR guidance (Fig. 3(c)). During training, the LDM backbone and the SR branch of TS-ControlNet are frozen, and the text branch is updated, enabling targeted text restoration while reducing disruption to overall image quality.

Given an HR target image $\mathbf{I}_{\text{HR}}$, we obtain the clean target latent $z_0 = \text{Enc}(\mathbf{I}_{\text{HR}})$ via the VAE encoder. We then sample a timestep $t \sim \mathcal{U}\{1, \ldots, T\}$ and noise $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, and construct the noised latent using the standard DDPM forward process Ho et al. (2020):

$$z_t \;=\; \sqrt{\bar{\alpha}_t}\, z_0 \;+\; \sqrt{1-\bar{\alpha}_t}\, \varepsilon, \qquad \bar{\alpha}_t = \textstyle\prod_{s=1}^{t}(1-\beta_s).$$

The diffusion model $\mathcal{D}_\theta$ predicts the noise residual conditioned on two control streams: (i) $\mathcal{C}_{\text{SR}}$, a spatial condition from a frozen SR-ControlNet that guides the overall structure based on the LR input image $\mathbf{I}_{\text{LR}}$; and (ii) $\mathcal{C}_{\text{TXT}}$, a textual condition from a trainable Text-ControlNet that guides text restoration using OCR-derived text-position pairs $\mathcal{S}_{\text{TXT}}$. To enable targeted text restoration, we fine-tune the text-control branch while keeping the main diffusion backbone and SR branch frozen; the OCR/VLM guider is adapted with a lightweight LoRA module.

At inference, we start from an initial Gaussian latent and use the EDM sampler Karras et al. (2022) with the same conditions to obtain the HR latent, which is then decoded into the image domain.

**Diffusion Loss with Residual Injection.** The frozen SR-ControlNet and the trainable Text-ControlNet produce residual hierarchies that are injected into the LDM backbone. Let $\mathbf{c}_t = \{c_t^\ell\}_{\ell=1}^{L}$ denote the set of residuals injected at $L$ U-Net blocks. We first encode the image-oriented and text-oriented prompts as

$$e_{\text{IMG}} = \psi([\mathcal{S}_{\text{IMG}}; P]), \qquad e_{\text{TXT}} = \psi([\mathcal{S}_{\text{TXT}}; P]),$$

where $\psi(\cdot)$ is the frozen CLIP text encoder and $[\cdot;\cdot]$ denotes prompt concatenation. For the $\ell$-th injection block, we blend the SR and text residuals as

$$c_t^\ell(\phi) = s_{\text{CTRL}}\Big[(1-\lambda_t)\mathcal{C}_{\text{TXT}}^\ell(z_t; e_{\text{TXT}}(\phi)) + \lambda_t \mathcal{C}_{\text{SR}}^\ell(z_t, f_{\text{LR}}; e_{\text{IMG}})\Big], \tag{1}$$

where $s_{\text{CTRL}}$ is a global residual scaling factor. The coefficient $\lambda_t$ is the image-guidance weight: $\lambda_t = 0$ yields text-centric guidance, $\lambda_t = 1$ yields image-centric guidance, and $\lambda_t = 0.5$ corresponds to uniform residual averaging.

To formalize the optimization target, let $\theta_0$ denote the frozen parameters (including the LDM backbone, CLIP text encoder, and SR-ControlNet), and let $\phi$ represent the trainable parameters, which encompass the Text-ControlNet and the VLM LoRA adapter. The diffusion backbone $\mathcal{D}_{\theta_0}$ predicts the noise residual conditioned on the mixed residual hierarchy $\mathbf{c}_t(\phi)$. We optimize only the parameter set $\phi$ using the following $\varepsilon$-prediction objective:

$$\min_{\phi} \mathcal{L}_{\text{text}} = \mathbb{E}_{z_0,\, t,\, \varepsilon}\left[\left\|\varepsilon - \mathcal{D}_{\theta_0}(z_t, t, \mathbf{c}_t(\phi))\right\|_2^2\right]. \tag{2}$$

**Synthetic Fine-Tuning Dataset.** To disentangle text legibility from holistic perceptual quality, we factorize synthetic data along two axes: global image quality, $\star \in \{\text{HQ}, \text{LQ}\}$, and glyph fidelity, $\diamond \in \{\text{pos}, \text{neg}\}$. Here, pos denotes correctly rendered text, whereas neg denotes corrupted or incorrectly rendered text.

For each text instance, we use the positive high-quality image $\mathbf{I}_{\text{HQ}}^{\text{pos}}$ as the canonical restoration target.

The remaining variants $\{\mathbf{I}_{\text{LQ}}^{\text{pos}}, \mathbf{I}_{\text{HQ}}^{\text{neg}}, \mathbf{I}_{\text{LQ}}^{\text{neg}}\}$ serve as controlled conditioning inputs that expose the model to global degradation, glyph corruption, or both. Thus, the model is always optimized toward the canonical text content and layout, rather than toward corrupted glyphs. Fig. 3(b) illustrates these four synthetic subsets and their matched guide prompts:

- **Positive-Text / High-Quality ($P_{\text{HQ}}^{\text{pos}}$).** High image quality with preserved character outlines and accurate text positioning.
- **Negative-Text / High-Quality ($P_{\text{HQ}}^{\text{neg}}$).** High image quality with intentionally corrupted character outlines while preserving the overall scene quality.
- **Positive-Text / Low-Quality ($P_{\text{LQ}}^{\text{pos}}$).** Low image quality with preserved character outlines and accurate text positioning.
- **Negative-Text / Low-Quality ($P_{\text{LQ}}^{\text{neg}}$).** Low image quality with intentionally corrupted character outlines and perturbed text positioning.

Each sample is represented by a composite conditioning tuple:

$$\mathcal{T}_{\star}^{\diamond} = \Big( \underbrace{z_{\star}^{\diamond}}_{\text{image latent}}, \underbrace{\psi(\mathcal{S}_{\text{IMG}})}_{\text{scene caption}}, \underbrace{\psi(\mathcal{S}_{\text{TXT}})}_{\text{text cues}}, \underbrace{\psi(P_{\star}^{\diamond})}_{\text{guide prompt}} \Big), \qquad \diamond \in \{\text{pos}, \text{neg}\},\ \star \in \{\text{HQ}, \text{LQ}\}. \tag{3}$$

Here, $z_{\star}^{\diamond} = \text{Enc}(\mathbf{I}_{\star}^{\diamond})$ is the first-stage latent of the synthetic image $\mathbf{I}_{\star}^{\diamond}$, and $\psi(\cdot)$ denotes the frozen CLIP text encoder. The text cue $\mathcal{S}_{\text{TXT}}$ is obtained by verbalizing the OCR-derived text-position pairs:

$$\mathcal{S}_{\text{TXT}} = \text{Verbalize}\left(\{(\mathcal{S}_{\text{text}}^{k}, \mathcal{S}_{\text{pos}}^{k})\}_{k=1}^{K}\right).$$

For all four variants generated from the same underlying text instance, the text-position pairs are extracted from the positive-text, high-quality variant. This provides a canonical reference for the correct text content and layout, allowing corrupted-text variants to be identified relative to the intended glyph structure.

### 3.2 Text–Image Balancing Scheduler

Although *TS-ControlNet* injects glyph-centric features, the allocation of text-oriented and image-oriented guidance along the diffusion trajectory is critical. We therefore introduce a scheduler that dynamically reweights the two guidance streams with a time-dependent coefficient. Let $i \in \{0, \ldots, N-1\}$ denote the reverse sampling index, where sampling proceeds from a high noise level $\sigma_i$ to a lower noise level $\sigma_{i+1}$. We define $\lambda_i \in [0, 1]$ as the image-guidance weight: $\lambda_i = 0$ corresponds to text-centric guidance, whereas $\lambda_i = 1$ corresponds to image-centric guidance.

**Mixed Embedding and Residual Guidance.** We first verbalize the OCR-derived text-position pairs as a text-level prompt:

$$\mathcal{S}_{\text{TXT}} = \text{Verbalize}\left(\{(\mathcal{S}^k_{\text{text}}, \mathcal{S}^k_{\text{pos}})\}_{k=1}^{K}\right).$$

The scene-level and text-level prompts are encoded separately and projected into a shared embedding space:

$$\begin{aligned} e_{\text{img}} &= W_{\text{img}}\, \psi(\mathcal{S}_{\text{IMG}}), \\ e_{\text{txt}} &= W_{\text{txt}}\, \psi(\mathcal{S}_{\text{TXT}}), \\ e^{\,i} &= (1-\lambda_i)\, e_{\text{txt}} + \lambda_i\, e_{\text{img}}, \end{aligned} \tag{4}$$

where $\psi(\cdot)$ is the frozen text encoder, and $W_{\text{img}}$ and $W_{\text{txt}}$ are projection layers to a shared conditioning space.

The same coefficient also modulates the residual injection. Let $r^{\text{TXT}}_{\ell}(i)$ and $r^{\text{SR}}_{\ell}(i)$ denote the residuals from the text and SR branches at injection block $\ell$, respectively. We compute the mixed residual as

$$\tilde{r}_{\ell}(i) = s_{\text{CTRL}}\left[(1-\lambda_i)\, r^{\text{TXT}}_{\ell}(i) + \lambda_i\, r^{\text{SR}}_{\ell}(i)\right], \tag{5}$$

where $s_{\text{CTRL}}$ is a global control scale. The mixed residual hierarchy is denoted by $\tilde{\mathbf{r}}(i) = \{\tilde{r}_{\ell}(i)\}_{\ell=1}^{L}$.

**Step Update with Mixed Guidance.** Given the mixed embedding $e^{\,i}$ and mixed residual hierarchy $\tilde{\mathbf{r}}(i)$, we compute the classifier-free guided noise estimate as

$$\widehat{\epsilon}_i = (1+\omega)\, \mathcal{D}_{\theta_0}\left(z^{(i)}, \sigma_i, e^{\,i}, \tilde{\mathbf{r}}(i)\right) - \omega\, \mathcal{D}_{\theta_0}\left(z^{(i)}, \sigma_i, \varnothing, \tilde{\mathbf{r}}(i)\right), \tag{6}$$

where $\omega$ is the guidance scale and $\varnothing$ denotes the empty prompt condition. Here, the negative branch removes only the text prompt while retaining the same control residuals $\tilde{\mathbf{r}}(i)$. Thus, our classifier-free guidance (CFG) performs prompt-level guidance under fixed structural controls rather than a fully unconditional generation, serving as an "empty-text conditional with controls" formulation. The latent is then updated by the EDM solver:

$$z^{(i+1)} = \text{EDMStep}\left(z^{(i)}, \widehat{\epsilon}_i, \sigma_i, \sigma_{i+1}\right). \tag{7}$$

At inference, we initialize $z^{(0)} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2\mathbf{I})$ and apply the above update over $N$ reverse sampling steps.

**Binary Ping–Pong Policy.** Empirically, we find that a binary alternating schedule is effective:

$$\lambda_i = \begin{cases} 0, & \text{if } \left\lfloor \frac{i+i_0}{\tau} \right\rfloor \bmod 2 = 0, \\ 1, & \text{otherwise}, \end{cases} \tag{8}$$

where $\tau \in \mathbb{N}$ is the toggle period and $i_0$ is an optional offset that determines the initial phase. Text-centric phases provide glyph-specific correction, while image-centric phases reinforce global structure and perceptual consistency. We also evaluate continuous schedules $\lambda_i = g(\sigma_i)$, including monotonic noise-level schedules, but the proposed binary ping–pong policy achieves the best OCR $F_1$ at comparable perceptual quality, as shown in Appendix C.

## 4 Experiments

### 4.1 Experimental Setup

We evaluate our method along two axes: semantic text restoration and perceptual SR quality. We report OCR-based $\mathbf{F_1}$ scores Chng et al. (2019) to quantify semantic correctness. Perceptual image quality is measured by no-reference IQA metrics, including MANIQA Yang et al. (2022), CLIP-IQA Wang et al. (2023), and MUSIQ Ke et al. (2021) (see Sec. A.1) [3]. Since LLaVA-NeXT Liu et al. (2024) is involved in our data-generation and guidance pipelines, we primarily rely on OpenOCR and GOT-OCR as independent OCR evaluators to ensure an unbiased evaluation. For the LLaVA-NeXT-based evaluation, we strictly employ

[3]All OCR and IQA scores are reported on a 0–100 scale, and higher is better.

Table 1: Quantitative comparison of OCR $F_1$ scores and perceptual image-quality metrics across datasets and models. All scores are reported on a 0–100 scale, and higher is better. Red and blue indicate the best and second-best scores, respectively. The LR input is included as a reference baseline. The bottom rows summarize the number of first- and second-place entries among SR methods; ties are counted for all tied methods.

| Dataset | Model | OpenOCR | | GOT-OCR | | LLaVA-NeXT | | MANIQA | | CLIP-IQA | | MUSIQ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | ×4 | ×8 | ×4 | ×8 | ×4 | ×8 | ×4 | ×8 | ×4 | ×8 | ×4 | ×8 |
| SVT | LR | 48.65 | 8.49 | 66.89 | 27.78 | 70.08 | 42.79 | 20.45 | 19.81 | 17.06 | 44.07 | 26.31 | 22.96 |
| | BSRGAN | 53.96 | 14.61 | 58.66 | 13.12 | 68.50 | 25.56 | 38.16 | 37.14 | 39.63 | 37.58 | 66.25 | 62.83 |
| | DiffBIR | 38.73 | 16.70 | 42.33 | 18.55 | 45.19 | 22.32 | 47.82 | 45.54 | 58.66 | 53.20 | 71.18 | 64.11 |
| | DiffTSR | 19.35 | 10.28 | 22.51 | 10.72 | 29.23 | 15.87 | 21.34 | 21.39 | 27.69 | 26.39 | 46.24 | 43.96 |
| | InvSR | 57.79 | 17.12 | 60.96 | 21.15 | 65.00 | 21.54 | 46.78 | 32.51 | 57.30 | 50.83 | 70.81 | 51.69 |
| | MARCONet | 0.00 | 0.00 | 0.25 | 0.00 | 0.00 | 0.00 | 30.92 | 30.84 | 24.43 | 24.76 | 27.26 | 27.02 |
| | MARCONet++ | 50.05 | 10.28 | 59.88 | 13.97 | 65.90 | 22.32 | 29.31 | 20.97 | 19.82 | 8.44 | 49.20 | 38.06 |
| | PiSA-SR | 63.30 | 17.53 | 65.23 | 24.05 | 67.75 | 37.76 | 37.41 | 34.02 | 44.30 | 18.39 | 61.87 | 30.24 |
| | Real-ESRGAN | 59.15 | 17.73 | 67.32 | 23.29 | 72.53 | 30.83 | 31.16 | 28.38 | 28.58 | 17.86 | 51.14 | 43.01 |
| | SwinIR | 54.61 | 14.61 | 63.53 | 20.75 | 73.03 | 30.48 | 26.32 | 22.05 | 44.50 | 26.68 | 34.55 | 30.33 |
| | StableSR | 59.88 | 20.95 | 63.76 | 24.43 | 73.91 | 43.24 | 24.75 | 23.16 | 32.18 | 23.38 | 24.44 | 16.22 |
| | SUPIR | 58.41 | 33.61 | 61.90 | 35.96 | 62.14 | 36.78 | 42.36 | 40.17 | 48.42 | 45.06 | 67.55 | 65.20 |
| | TAIR | 27.23 | 21.54 | 30.13 | 23.48 | 32.58 | 26.68 | 31.99 | 31.99 | 29.12 | 29.49 | 54.34 | 54.27 |
| | **GLYPH-SR** | **67.54** | **48.79** | **71.72** | **56.16** | **73.22** | **58.54** | **47.75** | **47.40** | **59.40** | **56.78** | **70.99** | **69.93** |
| SCUT-CTW1500 | LR | 14.63 | 0.53 | 23.55 | 4.76 | 47.23 | 10.18 | 28.92 | 16.39 | 31.16 | 26.19 | 25.82 | 17.71 |
| | BSRGAN | 24.67 | 3.37 | 21.86 | 3.54 | 35.10 | 3.88 | 51.41 | 46.21 | 47.44 | 37.83 | 67.52 | 66.05 |
| | DiffBIR | 24.71 | 4.76 | 23.82 | 5.10 | 30.71 | 4.64 | 62.37 | 54.75 | 61.90 | 49.89 | 71.19 | 63.16 |
| | DiffTSR | 19.77 | 2.95 | 15.98 | 2.86 | 23.69 | 2.90 | 35.39 | 35.49 | 30.59 | 31.88 | 55.83 | 50.43 |
| | InvSR | 29.57 | 2.09 | 26.41 | 2.17 | 34.50 | 2.43 | 57.75 | 29.65 | 55.94 | 29.62 | 69.25 | 40.29 |
| | MARCONet | 0.13 | 0.13 | 0.57 | 0.61 | 0.22 | 0.26 | 33.34 | 33.56 | 16.54 | 16.20 | 28.78 | 28.95 |
| | MARCONet++ | 22.72 | 2.35 | 20.63 | 2.60 | 33.10 | 2.60 | 34.65 | 14.75 | 19.58 | 8.06 | 43.61 | 30.27 |
| | PiSA-SR | 37.46 | 7.61 | 34.14 | 6.92 | 44.11 | 9.43 | 56.31 | 41.77 | 53.05 | 36.75 | 68.19 | 58.95 |
| | Real-ESRGAN | 31.31 | 5.02 | 26.94 | 5.64 | 43.25 | 7.74 | 40.81 | 28.37 | 43.43 | 20.95 | 52.66 | 39.99 |
| | SwinIR | 23.10 | 3.67 | 23.21 | 4.68 | 39.44 | 5.27 | 33.85 | 19.00 | 46.07 | 24.14 | 39.36 | 25.64 |
| | StableSR | 25.55 | 3.33 | 19.95 | 4.43 | 45.86 | 7.49 | 31.04 | 20.93 | 43.61 | 20.92 | 24.92 | 16.62 |
| | SUPIR | 18.26 | 5.43 | 17.61 | 6.26 | 24.37 | 7.00 | 57.35 | 55.46 | 51.68 | 47.02 | 66.96 | 65.55 |
| | TAIR | 33.98 | 10.74 | 29.44 | 9.23 | 41.67 | 12.14 | 65.38 | 63.60 | 47.05 | 36.57 | 67.08 | 66.38 |
| | **GLYPH-SR** | **38.26** | **11.09** | **36.96** | **14.71** | **42.90** | **14.67** | **70.33** | **61.94** | **57.88** | **48.21** | **70.31** | **63.43** |
| CUTE80 | LR | 65.80 | 39.38 | 50.58 | 36.31 | 80.47 | 67.18 | 28.93 | 17.29 | 36.80 | 22.58 | 37.64 | 17.32 |
| | BSRGAN | 73.09 | 55.21 | 56.02 | 46.57 | 83.97 | 71.18 | 44.22 | 42.07 | 55.73 | 54.31 | 69.13 | 67.33 |
| | DiffBIR | 68.88 | 59.56 | 48.82 | 44.71 | 81.84 | 70.53 | 51.04 | 47.53 | 72.64 | 62.09 | 69.06 | 64.62 |
| | DiffTSR | 61.08 | 54.39 | 47.48 | 42.33 | 73.71 | 63.30 | 33.94 | 33.55 | 38.47 | 42.95 | 58.74 | 57.46 |
| | InvSR | 72.46 | 56.42 | 55.62 | 45.18 | 84.75 | 72.46 | 50.30 | 37.66 | 67.78 | 62.43 | 70.66 | 57.69 |
| | MARCONet | 2.31 | 3.07 | 4.56 | 4.56 | 3.82 | 3.82 | 33.58 | 33.66 | 26.69 | 26.69 | 31.06 | 30.88 |
| | MARCONet++ | 69.21 | 50.58 | 54.39 | 45.18 | 81.02 | 71.50 | 31.88 | 21.03 | 34.90 | 22.72 | 54.15 | 44.24 |
| | PiSA-SR | 72.77 | 52.72 | 54.80 | 42.33 | 82.65 | 75.24 | 45.82 | 30.71 | 61.81 | 30.80 | 66.18 | 45.16 |
| | Real-ESRGAN | 73.71 | 59.18 | 58.79 | 49.27 | 84.23 | 74.33 | 38.20 | 35.17 | 48.71 | 36.46 | 60.65 | 56.55 |
| | SwinIR | 73.71 | 52.30 | 55.62 | 45.18 | 82.92 | 72.46 | 31.87 | 22.72 | 59.32 | 40.40 | 47.94 | 39.44 |
| | StableSR | 72.14 | 57.81 | 57.22 | 45.18 | 82.92 | 73.87 | 36.26 | 26.00 | 49.74 | 40.42 | 60.09 | 34.48 |
| | SUPIR | 70.85 | 58.01 | 51.87 | 42.81 | 82.11 | 70.20 | 47.50 | 46.38 | 62.62 | 61.67 | 68.26 | 67.04 |
| | TAIR | 55.21 | 42.81 | 43.77 | 40.87 | 69.87 | 62.20 | 58.25 | 37.11 | 49.76 | 36.84 | 72.06 | 55.06 |
| | **GLYPH-SR** | **73.09** | **63.66** | **55.62** | **45.65** | **85.01** | **73.71** | **49.77** | **47.75** | **65.93** | **65.85** | **69.96** | **68.85** |

| Model | GLYPH-SR | DiffBIR | TAIR | Real-ESRGAN | PiSA-SR | StableSR | SwinIR | BSRGAN | InvSR | SUPIR | DiffTSR | MARCONet | MARCONet++ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1st Rank | **20** | 6 | 4 | 3 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2nd Rank | **8** | 6 | 4 | 2 | 3 | 3 | 0 | 4 | 4 | 3 | 0 | 0 | 0 |

the original, publicly available weights rather than our low-resolution (LR) adapted model, treating it as an auxiliary robustness check. Experiments are conducted on three representative scene text benchmarks (details in Sec. B.2.1): SCUT-CTW1500 Liu et al. (2019), CUTE80 Risnumawan et al. (2014), and SVT Wang et al. (2011). We adopt *Juggernaut-XL* as the frozen LDM backbone; during training, we fine-tune only the proposed text-control branch and a lightweight LoRA adapter for the VLM guidance, while keeping both LDM and VLM base models frozen. Our synthetic corpus is generated using LLaVA-NeXT Liu et al. (2024), Nunchaku Cruanes et al. (2016), and SUPIR Yu et al. (2024b). Full data-generation pipelines, hyperparameters, and implementation details are provided in Appendix B.

Figure 4: Visualizing the simultaneous achievement of the bi-objective: high-fidelity text restoration and authentic global texture preservation.

### 4.2 Evaluation Results

As shown in Table 1, existing SR methods often show an imbalance between text fidelity and perceptual image quality. DiffBIR obtains strong no-reference IQA scores in several settings, but its OCR $F_1$ scores can be substantially lower on text-heavy benchmarks, suggesting that perceptual realism alone does not ensure character-level correctness. Conversely, methods such as StableSR can preserve OCR-friendly low-frequency structures in some cases, but often yield lower perceptual quality, indicating conservative restoration. Text-focused baselines such as DiffTSR, MARCONet, and MARCONet++ do not obtain top-ranked entries in this full-scene benchmark, suggesting that local or crop-oriented text restoration does not directly solve scene-level text SR.

GLYPH-SR achieves the largest number of top-ranked results across the evaluated OCR and perceptual metrics. In the rank summary, GLYPH-SR obtains 20 first-place and 8 second-place entries, while the nearest competitor, DiffBIR, obtains 6 first-place entries. The advantage is especially clear under severe degradation. On SVT $\times 8$, GLYPH-SR improves OpenOCR $F_1$ from 33.61 to 48.79, yielding a 15.18 percentage-point gain over the strongest baseline, while also achieving the best MANIQA, CLIP-IQA, and MUSIQ scores in the same setting. These results indicate that GLYPH-SR improves text readability without sacrificing perceptual image quality.

Fig. 4 illustrates the balance between glyph recovery and scene reconstruction. GLYPH-SR restores the text string "CARROLL STREET BAKERY" more coherently than the compared baselines, which either introduce spurious characters or leave the text blurred. The surrounding background is also preserved with fewer visible artifacts, such as mosaic patterns or unnatural cracks. This qualitative result supports the quantitative finding that text fidelity and perceptual quality should be evaluated jointly.

Fig. 5 illustrates the two failure modes. The left examples show glyph hallucination, where visually sharp strokes alter character identity and degrade OCR performance despite high IQA scores. The right examples show conservative restoration, where methods retain recognizable low-frequency text structures but provide limited SR improvement, resulting in blurry or low-contrast outputs. By preserving glyph topology while restoring realistic textures, GLYPH-SR reduces both failure modes.

Fig. 6 shows qualitative results under severe degradations from $\times 4$ to $\times 8$. Competing methods exhibit different failure modes. Diffusion-based methods such as DiffBIR can produce visually plausible but incorrect characters, while GAN-based methods such as BSRGAN often generate high-contrast but jagged or geometrically distorted glyphs. In contrast, GLYPH-SR better preserves text structure and legibility while maintaining plausible surrounding textures, consistent with the trade-off observed in Table 1.

Figure 5: Qualitative examples illustrating the trade-off between SR metrics (e.g., MANIQA, CLIP-IQA, MUSIQ) and OCR metrics ($F_1$, Accuracy) in scene text images. While some methods improve perceptual SR scores, they may degrade OCR performance, and vice versa.

Figure 6: Qualitative comparison with DiffBIR, Real-ESRGAN, BSRGAN, and GLYPH-SR on degraded LR inputs.

This confirms the trade-off between perceptual quality and OCR accuracy observed in Table 1. Methods that excel in one metric often fail in the other. GLYPH-SR provides a stronger balance between OCR accuracy and perceptual quality, particularly under the challenging $\times 8$ setting where many baselines degrade substantially.

### 4.2.1 Ablation Studies

Fig. 7 ablates the two OCR-derived guidance sources used by GLYPH-SR: (i) the recognized text string $\mathcal{S}_{\text{text}}$ and (ii) its spatial position cue $\mathcal{S}_{\text{pos}}$. We evaluate four settings: *full guidance*, *text-only*, *position-only*, and *no guidance*.

1. **Full guidance** ($\mathcal{S}_{\text{text}} + \mathcal{S}_{\text{pos}}$). Using both text and position cues yields the most faithful restoration, preserving stroke width, inter-letter spacing, and global text geometry.

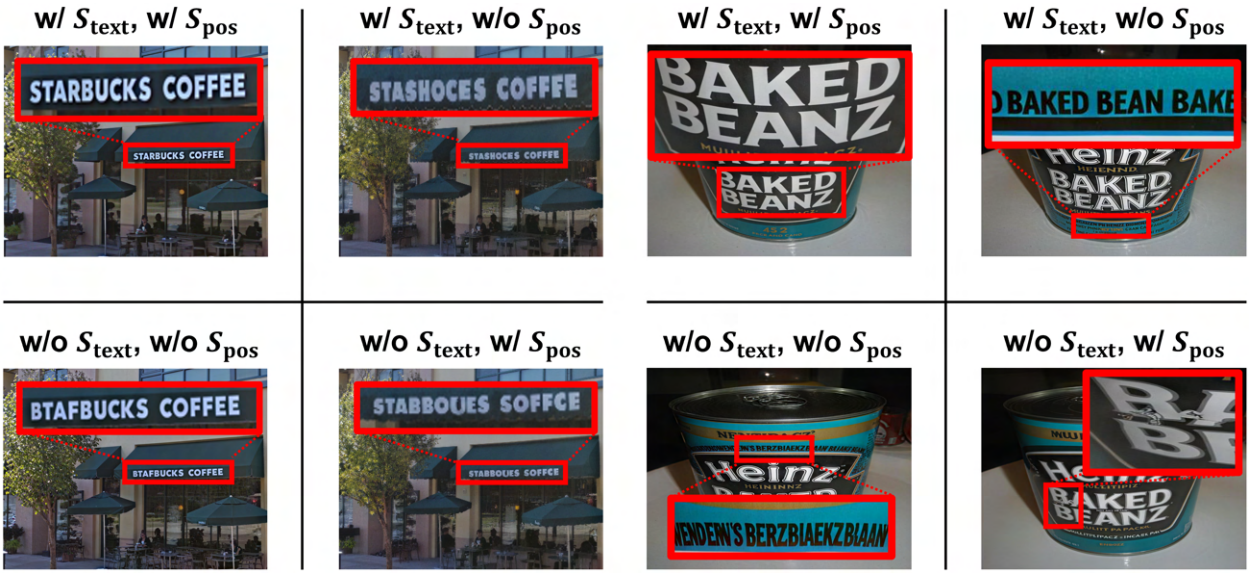



Figure 7: Prompt ablation using four combinations of OCR text strings ($\mathcal{S}_{\text{text}}$) and spatial position cues ($\mathcal{S}_{\text{pos}}$).

Figure 8: Qualitative comparison of DiffBIR and GLYPH-SR on non-Latin text, including Chinese and Devanagari-script examples.

2. **Text-only guidance** ($\mathcal{S}_{\text{text}}$ only). Removing positional guidance causes irregular kerning and warped baselines, e.g., "*STASHOES COFFEE*," indicating that text semantics alone cannot reliably anchor glyph layout.

3. **Position-only guidance** ($\mathcal{S}_{\text{pos}}$ only). Providing spatial cues without textual content preserves approximate text placement but produces partial or incorrect spellings, e.g., "*STABHOUES SOFFCE*." This suggests that location cues alone cannot resolve character-level ambiguity.

4. **No guidance**. Removing both guidance sources produces the most severe distortions, including hallucinated characters and unstable text geometry, similar to failures observed in weakly text-conditioned SR.

To examine cross-lingual generalization, we evaluate GLYPH-SR on non-Latin text from LSVT Sun et al. (2019) and IndicSTR Mathew et al. (2017). For this analysis, we replace the default English-centric guidance module with a multilingual VLM, Gemini 2.5 Flash, while keeping the diffusion backbone and TS-ControlNet unchanged. As shown in Fig. 8, GLYPH-SR better preserves complex glyph geometries and stroke patterns than DiffBIR on Chinese and Devanagari-script examples.

Figure 9: Qualitative Comparison of Curved Text SR (Curved Stress Test). Top: LR input. Middle: GLYPH-SR output. Bottom: magnified view.

To evaluate the robustness of our method under challenging conditions, we conducted stress tests focusing on curved and high-density text layouts. Fig. 9 presents a qualitative comparison on curved text, where the degree of curvature increases from left to right across the samples.

Figure 10: Qualitative Comparison of High-Density Text SR (Dense Stress Test). Top: LR input. Middle: GLYPH-SR output.

Furthermore, Fig. 10 highlights the capability of GLYPH-SR in highly dense and cluttered text environments. As the density and complexity of the text increase from left to right, GLYPH-SR consistently preserves individual characters and line structures better than the LR input. These qualitative examples suggest that GLYPH-SR can effectively manage complex text layouts, although broader quantitative evaluation remains necessary for a complete assessment.

## 5 Conclusion

Super-resolution research has traditionally prioritized perceptual image quality, often underemphasizing a critical requirement in text-rich scenes: legibility. This creates a gap in which models may produce visually sharp images whose text cannot be read correctly, as character-level errors are weakly captured by standard SR objectives and image-quality metrics. To address this issue, GLYPH-SR formulates scene-text SR as a bi-objective problem that jointly optimizes visual realism and text legibility.

We introduced a vision-language-guided diffusion framework with a dual-branch TS-ControlNet that integrates OCR-derived text and location cues with global scene guidance. We also proposed a factorized synthetic corpus for targeted text restoration and a dual-axis evaluation protocol that reports OCR $F_1$ together with perceptual IQA metrics. On SVT, SCUT-CTW1500, and CUTE80 at $\times 4$ and $\times 8$, GLYPH-SR improves OCR $F_1$ by up to 15.18 percentage points over strong baselines while maintaining competitive perceptual quality. Future work will extend quantitative multilingual evaluation, incorporate stronger geometric priors, and study tighter integration with end-to-end recognition systems.

## References


Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Jingye Chen, Haiyang Yu, Jianqi Ma, Bin Li, and Xiangyang Xue. Text gestalt: Stroke-aware scene text image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 285–293, 2022.

Zheng Chen, Yulun Zhang, Jinjin Gu, Xin Yuan, Linghe Kong, Guihai Chen, and Xiaokang Yang. Image super-resolution with text prompt diffusion. *arXiv preprint arXiv:2311.14282*, November 2023.

Chee Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, ChuanMing Fang, Shuaitao Zhang, Junyu Han, Errui Ding, et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1571–1576. IEEE, 2019.

Simon Cruanes, Jasmin Blanchette, and Andrei Popescu. Extending nunchaku to dependent type theory. In *Electronic Proceedings in Theoretical Computer Science (EPTCS)*, volume 210, pp. 3–12. Open Publishing Association, 2016.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):295–307, 2015.

Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S Ren, Radu Timofte, Yuan Gong, Shanshan Lao, Shuwei Shi, Jiahao Wang, Sidi Yang, et al. Ntire 2022 challenge on perceptual image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 951–967, 2022.

Junhao Gu, Peng-Tao Jiang, Hao Zhang, Mi Zhou, Jinwei Chen, Wenming Yang, and Bo Li. Improving consistency in diffusion models for image super-resolution. *arXiv preprint arXiv:2410.13807*, 2024. URL `https://arxiv.org/abs/2410.13807`.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 633–651. Springer, 2020.

Tero Karras, Miika Aittala, Samuli Laine, Timo Herva, and Jaakko Lehtinen. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022. URL `https://arxiv.org/abs/2206.00364`.

Jun Ke, Guy Hacohen, Phillip Isola, William T. Freeman, Michael Rubinstein, and Eli Shechtman. Musiq: Multi-scale image quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8827–8837, 2021.

Xiaoming Li, Wangmeng Zuo, and Chen Change Loy. Learning generative structure prior for blind text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10103–10113, June 2023.

Xiaoming Li, Wangmeng Zuo, and Chen Change Loy. Enhanced generative structure prior for chinese text image super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1833–1844, October 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *European Conference on Computer Vision*, pp. 430–448. Springer, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

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

Jianqi Ma et al. A text attention network for spatial deformation robust scene text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Minesh Mathew, Mohit Jain, and CV Jawahar. Benchmarking scene text recognition in devanagari, telugu and malayalam. In *Proceedings of the 6th International Workshop on Multilingual OCR*, pp. 1–6, 2017.

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, 2025. URL `https://arxiv.org/abs/2506.09993`.

Chihiro Noguchi, Shun Fukuda, and Masao Yamanaka. Scene text image super-resolution based on text-conditional diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1485–1495, 2024.

Anhar Risnumawan, Palaiahankote Shivakumara, Chee Seng Chan, and Chew Lim Tan. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, 41(18):8027–8048, 2014.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2333–2343, 2025.

Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Craig CT Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, et al. Icdar 2019 competition on large-scale street view text with partial labeling (rrc-lsvt). In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1557–1562, 2019.

Mingyu Sung et al. A novel vlm-guided diffusion model for remote sensing image super-resolution. *IEEE Geoscience and Remote Sensing Letters*, 22:LGRS–2025, 2025.

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

Jianyi Wang, Zhaoyi Wang, Xiangyu Zhang, Errui Ding, Hao Tang, and Ping Luo. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision (IJCV)*, 2024. URL `https://github.com/IceClear/StableSR`.

Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *2011 International conference on computer vision*, pp. 1457–1464. IEEE, 2011.

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 (ICCV) Workshops*, pp. 1905–1914, October 2021.

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.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2286–2295, 2022.

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

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

Zongsheng Yue, Kang Liao, and Chen Change Loy. Arbitrary-steps image super-resolution via diffusion inversion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23153–23163, 2025.

Kai Zhang et al. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

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 (CVPR)*, June 2018a.

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

# A Metric-Based Evaluation

## A.1 Evaluation Metrics

We evaluate our method along two primary axes: (1) the *semantic integrity* of textual content, and (2) the *perceptual quality* of the reconstructed images. Accordingly, we organize the metrics into two groups.

**OCR Metrics.** To assess text restoration, we report character-level precision, recall, $F_1$, accuracy, and normalized edit similarity:

$$F_1 = \frac{2PR}{P+R}, \qquad \text{1-NED} = 100 \times \left(1 - \frac{\text{ED}(y, \hat{y})}{\max(|y|, |\hat{y}|)}\right),$$

where $y$ and $\hat{y}$ denote the ground-truth and predicted text strings, respectively. Higher values are better.

**Image-Quality Metrics.** For image quality, we report full-reference metrics, including PSNR, SSIM, and LPIPS, and no-reference IQA metrics, including MANIQA, CLIP-IQA, and MUSIQ.

- **Peak Signal-to-Noise Ratio (PSNR)** (↑): log-scaled pixel-level similarity to the reference image.
- **Structural Similarity Index (SSIM)** (↑): evaluates luminance, contrast, and structural consistency in line with human perception, scaled to 0–100.
- **Learned Perceptual Image Patch Similarity (LPIPS Zhang et al. (2018a))** (↓): deep-feature distance reflecting perceptual differences, scaled to 0–100.
- **Multi-Dimension Attention Network for No-Reference IQA (MANIQA Yang et al. (2022))** (↑): no-reference quality score based on attention-driven features, scaled to 0–100.
- **CLIP-IQA** (↑): a no-reference image-quality metric based on CLIP representations, scaled to 0–100.
- **Multi-Scale Image Quality Transformer (MUSIQ Ke et al. (2021))** (↑): transformer-based no-reference IQA that aggregates multi-resolution cues.

## A.2 Misalignment Between Metrics and Human Perception

SR papers still default to **PSNR**, **SSIM**, and **LPIPS**. Although convenient, these scores can diverge from qualitative legibility and visual plausibility, especially when the LR input is heavily degraded. Fig. 11 offers four counter-examples that highlight three recurring failure modes.

**Perceptual vs. Semantic Fidelity.** In Figure 11, the "`HOMER BREWING COMPANY`" sign is reconstructed cleanly by GLYPH-SR yet receives lower PSNR and SSIM than Real-ESRGAN, whose output contains aliasing and hallucinated glyphs.

**Metrics Can Be Misleading.** Across multiple benchmarks we frequently observe visually superior outputs that score lower on PSNR/SSIM/LPIPS (see red vs. blue values in Figure 11). This misalignment—echoed by prior studies Blau & Michaeli (2018); Jinjin et al. (2020); Gu et al. (2022); Yu et al. (2024a)—underscores the danger of metric-only evaluation. For text-aware SR, qualitative inspection and OCR-aware metrics should complement conventional IQA metrics.

Figure 11: Each triplet shows (left) the input LR image, (middle) a strong baseline, and (right) GLYPH-SR. Despite GLYPH-SR producing visibly sharper text, its PSNR/SSIM/LPIPS scores (blue) are often lower than those of the baseline (red). These examples show that traditional full-reference metrics may not adequately reflect text legibility in scene-text SR.

# B Experiment details

## B.1 Reproducibility Statement

**Data Generation & Fine-Tuning Workflow**

1. **Stage 1 – Scene Description Extraction**
   `dataset_generater/make_dataset_get_desc.py`
   `./datasets/descriptions/` containing: *{id, image_path, ocr_text, caption}*.

2. **Stage 2 – Augmented Prompt Synthesis**
   `third_party/make_dataset_with_nunchaku/`
   `make_dataset_with_augmentation.py`
   The script invokes a Nunchaku-backed FLUX pipeline for efficient synthetic image generation. The output is a paired folder structure: `./datasets/aug/{hq, lq}`.

3. **Stage 3 – Negative/HQ Pairing**
   `dataset_generater/make_dataset_Neg_HQ.py`
   Generates explicit (LQ, HQ) pairs and the associated prompt metadata required by GLYPH-SR. Final training files are placed under `./datasets/final/`.

4. **Stage 4 – Fine-Tuning**
   `train_GLYPH_SR.py`

```
python3 train_GLYPH_SR.py \
        --data_root ./datasets/ \
        --cfg       GLYPH-SR/model_configs/model_config.yaml
```

## B.2 Setup

All experiments were conducted on a workstation equipped with three NVIDIA A6000 GPUs, all utilized concurrently for training and inference, an Intel Xeon W9-3475X CPU (36 cores, 72 threads), and 256GB RAM. The system runs Ubuntu 24.04.2 LTS and uses a 3.7TB NVMe SSD for storage. The models were implemented in PyTorch 2.5.1 with CUDA 11.8.

### B.2.1 Evaluation Datasets

**ICDAR2017 (International Conference on Document Analysis and Recognition).** The ICDAR2017 Robust Reading Challenge dataset consists of scene text images designed to test text detection and recognition systems under real-world conditions. It includes both high- and low-quality images that exhibit various degradations such as blur, LR, and noise. This diversity makes it well-suited for fine-tuning vision-language models to enhance OCR robustness. ICDAR2017 is used only for adapting the OCR/VLM guider and is not included in the final benchmark evaluation. We use the training split only and ensure that no evaluation images from SVT, SCUT-CTW1500, or CUTE80 are used during guider adaptation.

**SCUT-CTW1500 (Curved Text in the Wild).** SCUT-CTW1500 is a large-scale scene text detection benchmark featuring 1,500 images with over 10,000 annotated curved text instances. The dataset includes a wide variety of natural scenes such as street views, signboards, and shop names, with text appearing in arbitrary orientations, lengths, and curvature. It is especially known for its high diversity in text shape and layout, which makes it well-suited for evaluating the robustness of text detection and SR models in processing long and curved text lines. SCUT-CTW1500 is widely used for benchmarking models designed to process irregular and multi-oriented scene text under real-world conditions.

**CUTE80 (Curve Text).** CUTE80 is a compact yet challenging dataset containing 80 high-resolution images, specifically curated to evaluate curved text detection and recognition systems. The dataset features a range of naturally curved and perspective-distorted text instances embedded in complex backgrounds such as logos, signs, and posters. Despite its small size, CUTE80 is frequently used in literature to benchmark the generalization ability of text-focused models on non-horizontal and non-linear text structures. Its emphasis on difficult geometric deformations makes it a useful supplement to larger datasets for testing text-specific visual models under challenging conditions.

**SVT (Street View Text).** SVT is a benchmark dataset collected from Google Street View, consisting of 647 images with approximately 2,000 annotated text instances. It features naturally occurring scene text with various distortions, backgrounds, lighting conditions, and orientations. Despite its relatively small size, SVT is widely used in the literature for benchmarking the performance of OCR and text SR models under real-world conditions. Its challenging scenarios make it suitable for evaluating model generalization and robustness in unconstrained environments.

#### B.2.2 Pre-trained Models

**LLaVA-NeXT.** We employ **LLaVA-NeXT** Liu et al. (2024) as the vision–language front-end that extracts semantic context from LR inputs. LLaVA-NeXT couples a CLIP-ViT visual encoder with a 7-B parameter LLM and is instruction-tuned on a large multimodal corpus, yielding state-of-the-art performance in fine-grained grounding, captioning, and region-level reasoning. Within our pipeline it automatically produces (i) image-level captions (*IMG prompts*) and (ii) spatially aligned OCR strings (*OCR prompts*); both streams are fed as high-level conditions to the diffusion backbone.

**JuggernautXL (SDXL-based).** For image generation we adopt **JuggernautXL**, a publicly released checkpoint built on *SDXL-base 1.0* and further fine-tuned for improved sharpness and color fidelity. The underlying SDXL architecture is trained on billions of image–text pairs and natively supports $1024 \times 1024$ resolution.

#### B.2.3 Key Hyper-parameters

- **Vision–Language Encoder.** The base LLaVA-NeXT weights are frozen, and only a lightweight LoRA adapter is trained. Thus, the VLM contributes 5.9M trainable parameters while keeping the full backbone frozen.
- **First Stage (VAE).** A $256 \times 256$ auto-encoder (4 latent channels, $4\times$ down-sampling) maps RGB images to a $64 \times 64 \times 4$ latent grid.
- **Denoising and Sampling.** We use the standard 1 000-step DDPM schedule wrapped by Restore-EDM sampling (default: 50 inference steps, classifier-free guidance scale annealed from 7.5 to 4.0).

#### B.2.4 Synthetic Training Data

To train the TS-ControlNet, we curate a purpose-built synthetic dataset with four mutually exclusive partitions: *Positive/High-Quality*, *Positive/Low-Quality*, *Negative/High-Quality*, and *Negative/Low-Quality*. Each split is created by selectively degrading either global content or localized glyph regions while keeping spatial layout and annotations intact. This design lets the network disentangle text-specific cues from general image priors.

**Step 1: Prompt-Metadata Extraction.** Using a pretrained LLaVA-NeXT encoder and YAML-defined prompt templates, we batch-process scene text images and record three fields in `JSONL`: image `id`, OCR text, and a scene-level `prompt`. Figure 12 illustrates the resulting metadata, produced by `make_dataset_get_desc.py`.

Figure 12: Step 1: JSONL metadata—`id`, OCR text, and scene prompt—generated by LLaVA-NeXT.

Figure 13: Step 2: Prompt augmentation with stylistic keywords to boost visual diversity.

**Step 2: Stylistic Augmentation.** We prepend each prompt with a style token (e.g., *sunset glow*, *cinematic bokeh*) via `make_dataset_with_augmentation.py`. The enriched prompts drive a Flux-based diffusion model equipped with ControlNet and LoRA modules to generate visually diverse high-quality samples (Fig. 13).

**Step 3: Negative High-Quality Pairs.** The `make_dataset_Neg_HQ.py` script corrupts text regions at the glyph level while leaving global detail untouched, yielding hard negative examples. We generate four variants for each underlying text instance by independently controlling global image quality and glyph fidelity. Positive denotes correctly rendered text, whereas negative denotes intentionally corrupted glyphs. The positive-HQ variant provides the canonical text and layout annotation; the remaining variants are used to expose the model to controlled text and image degradations Yu et al. (2024b).

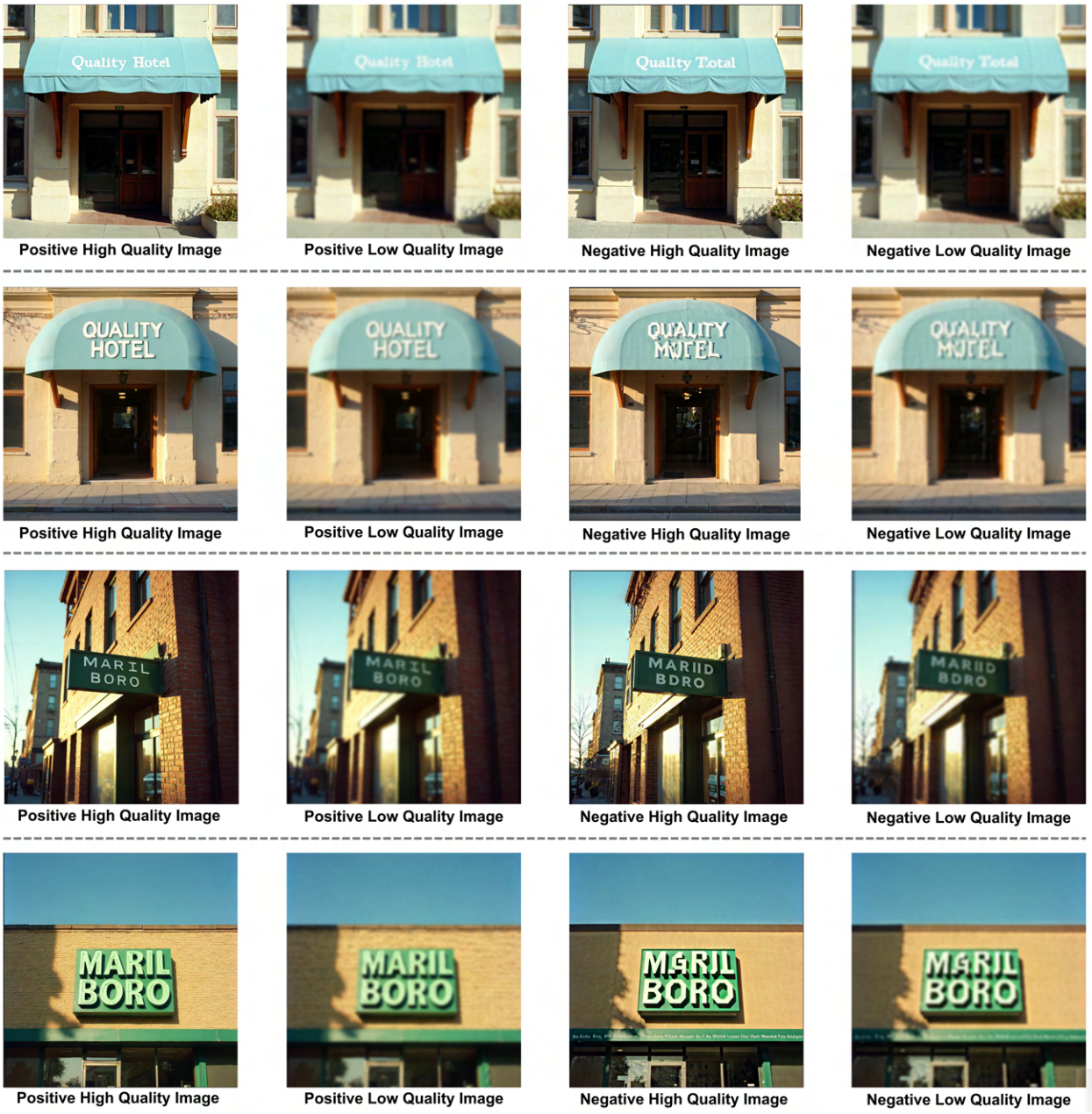



Figure 14: Step 4: Four synthetic subsets used for text-aware SR training.

**Step 4: Final Dataset Assembly.** All positive and negative images are merged into the four target splits shown in Fig. 14.

# C Experiment results

## C.1 Compare character generation to other models.

We evaluate the character-level restoration quality of each SR model using multiple OCR systems across degradation levels. The results in Table 2 show substantial improvements, especially under hard conditions. When LLaVA-NeXT is used as an OCR evaluator, we prompt it with: "Please perform OCR on this image." The same text-normalization rule is applied to all OCR outputs and ground-truth strings. Additionally, both predicted and ground truth texts were normalized by removing non-alphabetic characters and ignoring case sensitivity before computing the metrics.

Fig. 19 provides a qualitative comparison between GLYPH-SR and baselines at magnification factors of $\times 4$, $\times 8$, and an extreme $\times 16$. GLYPH-SR continuously reconstructs glyph outlines, stroke widths, and kerning while remaining true to the underlying truth, while harmonizing color and brightness with the surrounding background area. This visual evidence corroborates the quantitative gap observed in Table 1: models that optimize solely for perceptual metrics (e.g. DiffBIR, Real-ESRGAN) or for edge contrast fall short on OCR fidelity once the scale factor exceeds $\times 8$.

## C.2 Ablation study

In this ablation, we test an alternative to our binary ping-pong policy: several static **"Mixing"** policies. These policies use a constant mixing ratio $\lambda_t = C$ (where $C \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$) to blend the text and image guidance at every step, rather than alternating. The results in Table 4 demonstrate that this static mixing leads to an unstable and sub-optimal trade-off.

As detailed in Table 5, increasing $w$ generally improves perceptual metrics (e.g., CLIP-IQA, MUSIQ) by strengthening the generative prior. However, we observed a distinct trade-off where excessive guidance ($w > 7.5$) leads to a degradation in OCR performance.

The control scale modulates the influence of our Text-SR Fusion ControlNet. As presented in Table 6, we identified severe failure modes associated with high values of $s_{\text{CTRL}}$. When the scale exceeds 1.0, the strong injection of control features disrupts the natural image statistics of the diffusion backbone, leading to over-conditioning.

To investigate the trade-off mechanism empirically, we visualized the internal noise prediction maps $\epsilon_\theta$ across diffusion timesteps, as shown in Fig. 21. We compared the feature map evolution under two distinct conditions: a baseline with accurate text guidance versus a perturbed setting with high guidance error (OCR error). As observed in the heatmaps, the introduction of incorrect conditioning generates significant spatial anomalies, which we term 'prediction noise outliers' (highlighted in red boxes). Unlike the baseline, where noise predictions converge coherently, the high-error setting exhibits persistent instability throughout the intermediate steps (e.g., Step 25).

Based on these visual findings, we discuss the potential mechanism by which incorrect text conditioning affects the generation process:

- **Noise Dispersion:** When incorrect conditioning is injected, the predicted noise values fail to localize around the text regions. Instead, as shown in Fig. 21, high-variance noise artifacts become spatially dispersed across the entire domain.
- **Perturbation of Denoising Trajectory:** In diffusion models, the noise predictor $\epsilon_\theta(z_t, t, c)$ estimates a noise vector subject to a magnitude constraint at each timestep. When the semantic condition $c$ conflicts with the visual features (due to erroneous OCR), conflicting text conditioning may redistribute denoising activity away from localized glyph regions, potentially perturbing the denoising trajectory beyond text regions. However, this feature-map analysis is diagnostic rather than causal, and a more rigorous attribution study is left for future work.

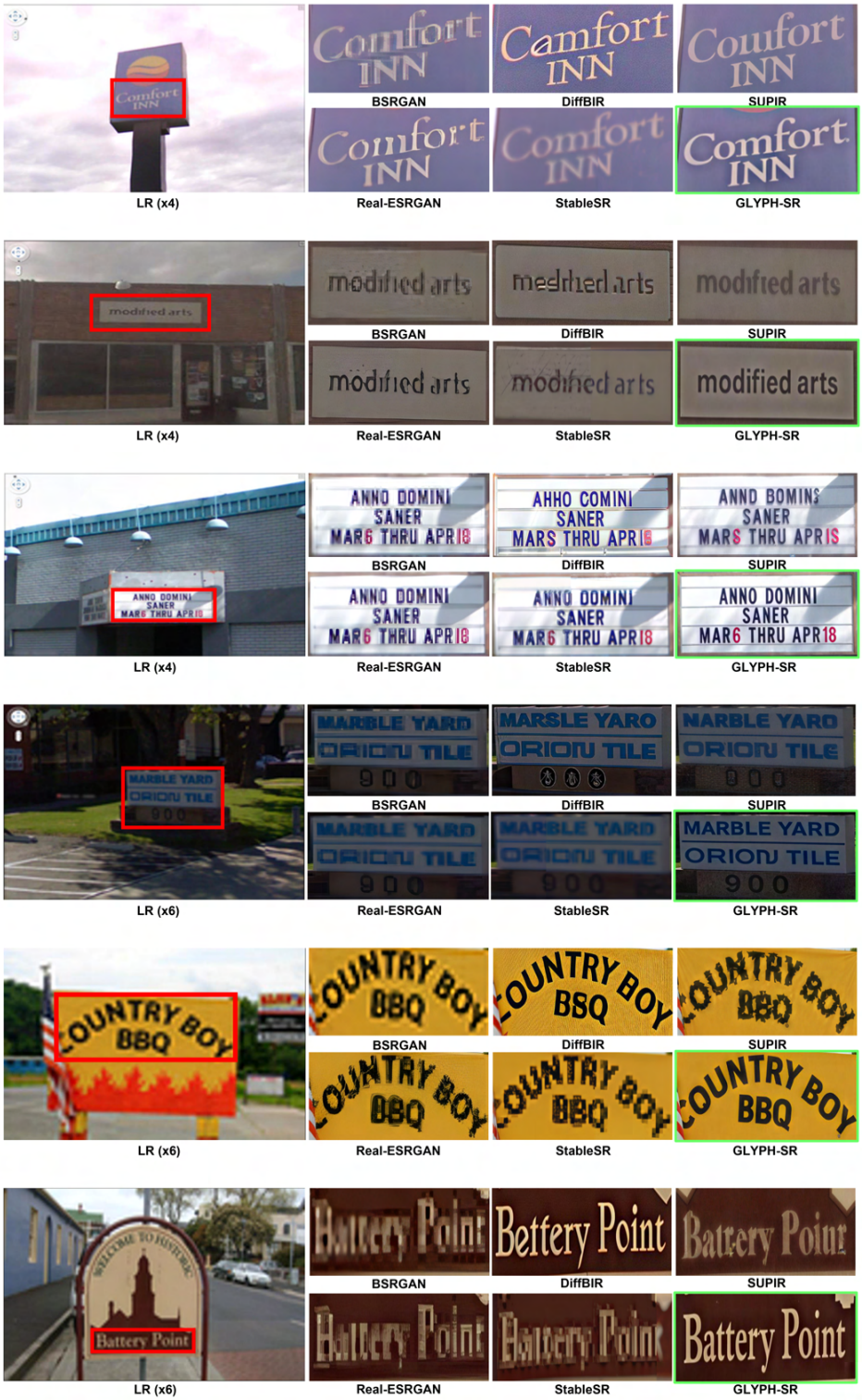



Figure 15: Qualitative comparison.

- **Influence on Surrounding Regions:** Consequently, we observe that incoherent noise updates driven by semantic conflict can introduce uncooperative artifacts in the surrounding background. These observations suggest that accurate text guidance appears to stabilize glyph restoration and

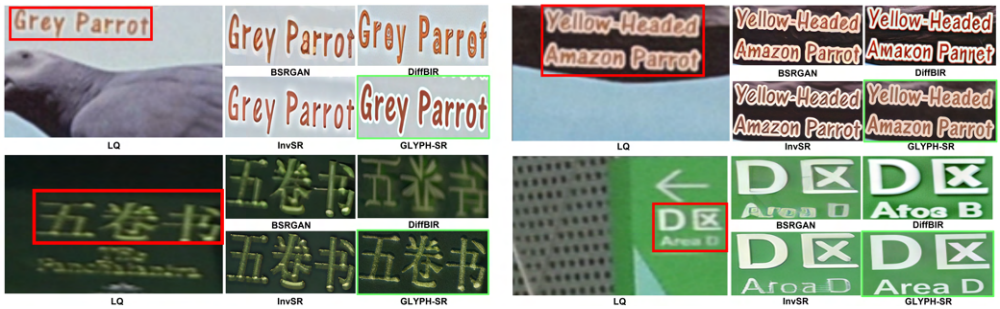



Figure 16: Qualitative comparison of text-focused SR on the Real-Text dataset. Each block shows a LQ input (left) and the outputs of four representative SR methods: BSRGAN, DiffBIR, InvSR, and the proposed GLYPH-SR (right). GLYPH-SR consistently reconstructs sharper glyph boundaries, preserves accurate character shapes, and restores correct textual semantics across diverse scenes—including license plates, bird labels, warning signs, Chinese characters, and directional symbols—while competing methods often exhibit distortions, blurring, or hallucinated characters.

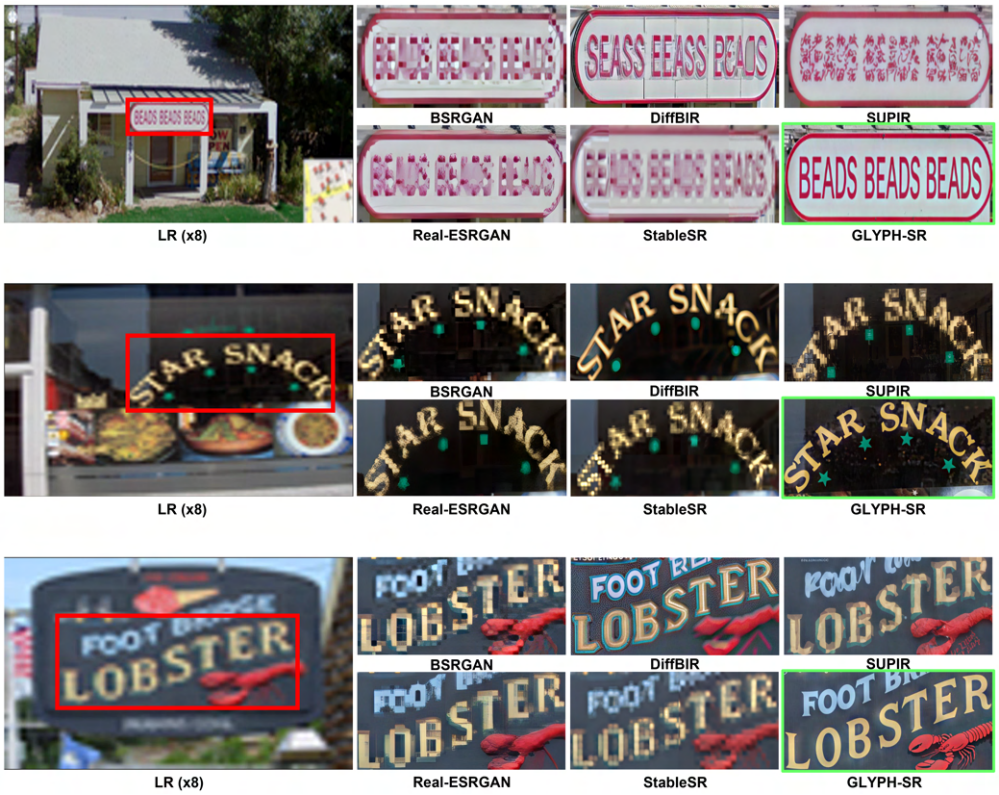



Figure 17: Qualitative comparison of scene text SR under various degradation scales (×4, ×6, ×8). While prior methods often blur or hallucinate characters, **GLYPH-SR** accurately restores readable, coherent text. Zoom in for detail.

may also help preserve surrounding textures, although the relationship between text fidelity and global image quality metrics is not strictly monotonic.

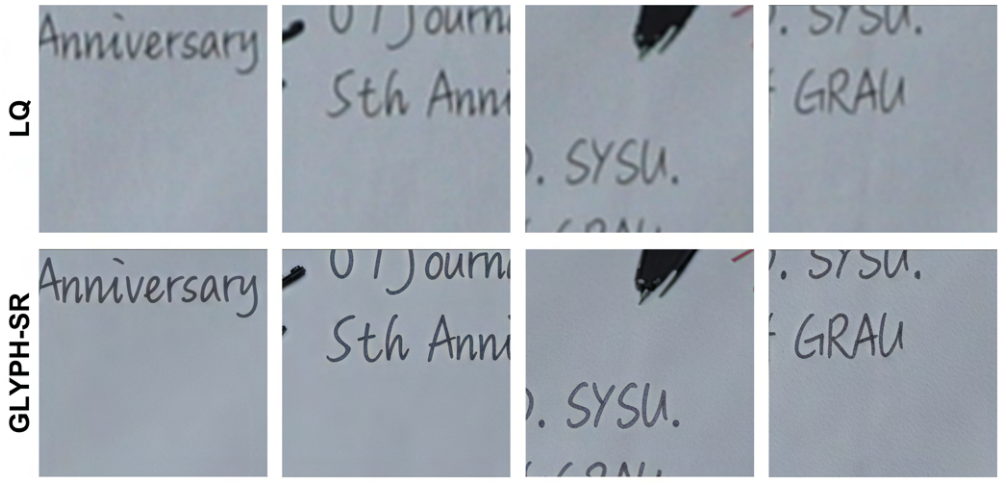



Figure 18: Qualitative Comparison of Handwriting SR. Visual comparison between the LQ input (top row) and the output generated by GLYPH-SR (bottom row) on various handwritten texts.

Figure 19: Qualitative comparison of text-centric SR results at ×4, ×8 and ×16 scales.

### C.3 Computational Footprint and Practical Efficiency

**Setup.** We benchmark inference on a 4× SR task with 512×512 inputs. Times are mean ± std. over repeated runs. For methods that require a large VLM (SUPIR and GLYPH-SR), we used two NVIDIA A6000 GPUs; reported peak VRAM is the sum across both devices.

GLYPH-SR trades extra parameters and memory for markedly better text fidelity: it couples a restoration backbone with a powerful OCR/VLM to reason about LR text. This design improves accuracy but introduces a computational bottleneck. To mitigate the cost while keeping readability gains, we will pursue:

- **Lighter VLM Guidance.** Replace the current general-purpose VLM with a compact, LR-text-specialized guider (or distill the guider), reducing parameter count and latency with minimal loss in guidance quality.

- **Inference Optimization ("Block Caching").** Cache and reuse guidance features that repeat across diffusion steps/tiles (e.g., projected text embeddings and cross-attention KV maps), skipping redundant compute and lowering end-to-end runtime.

Figure 20: Performance Comparison under Extreme Degradation. We compare SR results with extremely LR input images. Despite the inherent difficulty of achieving full restoration in these Extreme LR scenarios, GLYPH-SR demonstrates a unique capability for semantic reconstruction. As shown by the comparison against the LR input and the GT, GLYPH-SR successfully recovers partial text structure and generates sharper outlines, showcasing its effectiveness even when input quality is severely degraded.

These directions aim to preserve GLYPH-SR's strengths ("looks right and reads right") while improving deployability under realistic compute budgets.

Table 8 highlights a clear but controllable accuracy–efficiency trade-off. While GLYPH-SR employs a high-capacity LLaVA-NeXT-8B guider by default to handle severe degradations, our ablation study demonstrates that significantly lighter models can serve as efficient alternatives. Specifically, replacing the 8B guider with the 4B model (R-4B) reduces the model size by 50%, yet the OpenOCR $F_1$ score drops by a negligible 0.32 points (from 73.09 to 72.77). Even with the 3B model (PaliGemma-3B), which represents a ∼62.5% reduction in parameters, the model maintains robust text recognition capabilities with a moderate decline of 2.56 points in OpenOCR. These results suggest that smaller guiders can preserve much of the OCR performance. However, a complete efficiency comparison requires measuring latency and memory under the same inference setup.

We conducted a proof-of-concept experiment on the CUTE80 dataset under an extreme downscaling factor of ×16, a setting where severe degradation renders text nearly illegible and prone to detection failures. As summarized in Table 9, integrating the dual-stage restoration framework Sung et al. (2025) significantly improved OpenOCR performance (+4.14), demonstrating its effectiveness in mitigating hallucinations and recovering structural details despite detection errors. Furthermore, refining the inference process with Chain-of-Thought (CoT) prompting notably enhanced perceptual quality metrics (ManiQA and ClipIQA) while boosting recognition accuracy, confirming its capacity to effectively balance the trade-off between fidelity and perception. These preliminary results indicate that extreme ×16 degradation remains challenging. Additional modules can partially improve either OCR accuracy or perceptual quality, but the trade-off is not fully resolved.

**Trainable parameters.** Although the full model size is large due to the VLM, our fine-tuning recipe is lightweight. We freeze the diffusion backbone and update only two components:

1. **TS-ControlNet branch** (≈54.8M parameters) that handles text-guidance fusion.

2. **VLM LoRA adapter** (≈5.9M parameters) with low rank ($r$=8), `lora_alpha` of 32, and dropout of 0.05.

Figure 21: Intermediate Feature Analysis under OCR Error Rates.

To minimize memory further, the large frozen VLM is loaded in 4-bit quantization (`nf4` with double quantization via BitsAndBytes).

**Inference latency.** The OCR/VLM guider is the main overhead driver. Out of a total per-image latency of $38.25 \pm 1.28$ seconds (Sec. C.3), the VLM component accounts for $\approx 8.46$ seconds. Notably, while integrating the VLM increases total parameter count, the latency impact is not proportional. In practice, we retain training practicality with only 60.7M trainable parameters and observe that the rise in inference time is moderate relative to the parameter growth, yielding a favorable trade-off between accuracy (readability and IQA) and compute.

**Implication.** These results align with our compute study (Table 7): GLYPH-SR deliberately expends parameters on guidance quality to secure text fidelity, yet its fine-tuning footprint remains compact and deployable. Further efficiency gains are compatible with our design (e.g., lighter LR-text guiders and block caching for reusable guidance features).

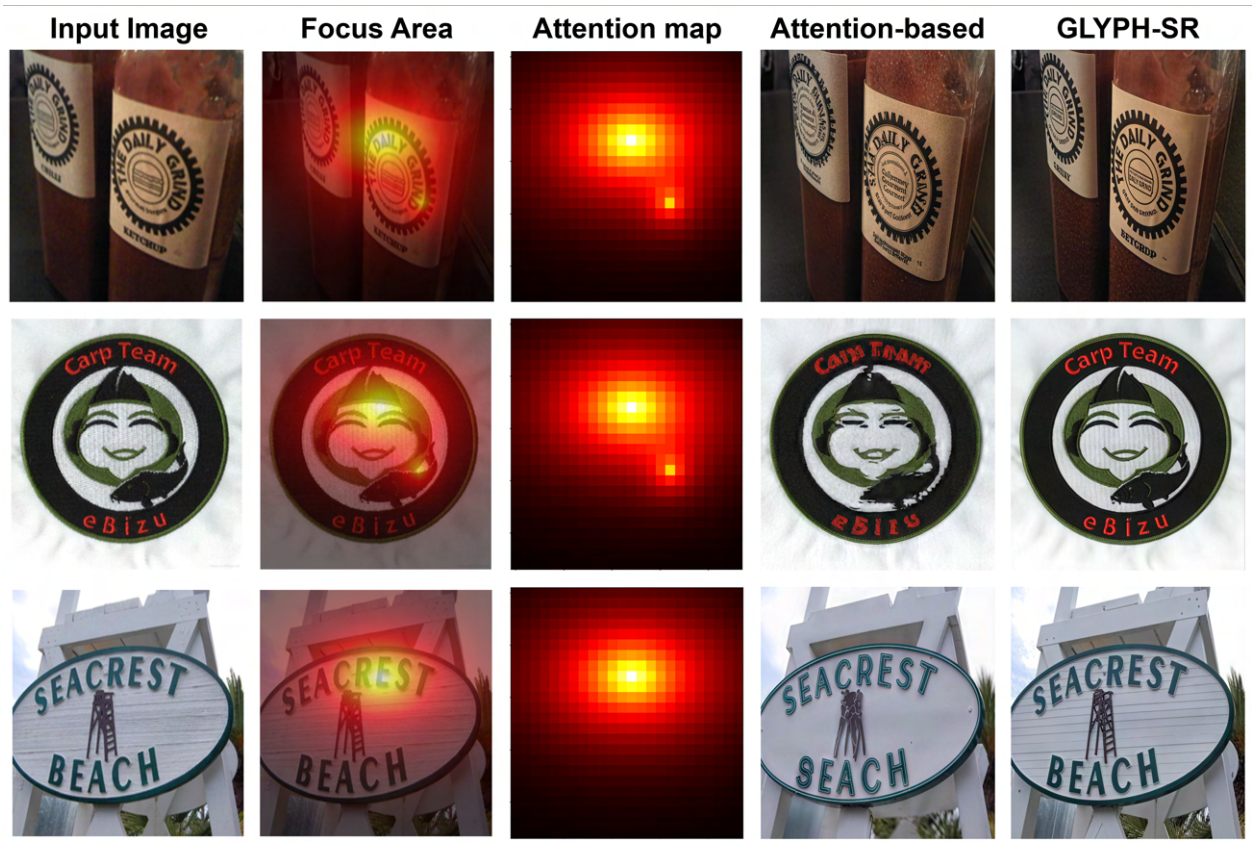



Figure 22: Comparison with explicit attention injection. We investigated an alternative design that explicitly injects OCR attention maps into the backbone. However, this explicit attention guidance proves detrimental; instead of refining features, it interferes with the backbone's pre-trained priors, causing the model to focus on irrelevant noise (as shown in the Attention Map). In contrast, our final GLYPH-SR (rightmost) harmonizes text guidance without such interference, demonstrating successful restoration.

Table 2: Quantitative comparison of OCR performance on images degraded by various factors and restored using six SR models, evaluated across three benchmark datasets and three OCR systems. Red and blue denote the best and second-best results, respectively.

| | OpenOCR | | | | | GOT-OCR | | | | | LLaVA-NeXT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | $F_1$ score | 1-NED | Accuracy | Precision | Recall | $F_1$ score | 1-NED | Accuracy | Precision | Recall | $F_1$ score | 1-NED | Accuracy |
| **SVT(x4)** | | | | | | | | | | | | | | | |
| LR | 61.41 | 40.28 | 48.65 | 26.37 | 32.14 | 75.70 | 59.91 | 66.89 | 34.60 | 50.25 | 83.27 | 60.50 | 70.08 | 13.50 | 53.94 |
| HR | 74.73 | 87.86 | 80.76 | 27.69 | 67.73 | 81.18 | 82.97 | 82.06 | 34.80 | 69.58 | 87.68 | 82.14 | 84.82 | 17.77 | 73.65 |
| BSRGAN | 57.03 | 51.19 | 53.96 | 28.04 | 36.95 | 69.63 | 50.68 | 58.66 | 31.29 | 41.50 | **84.60** | 57.55 | 68.50 | 15.82 | 52.09 |
| DiffBIR | 41.49 | 36.31 | 38.73 | 20.21 | 24.01 | 48.34 | 37.65 | 42.33 | 24.47 | 26.85 | 63.20 | 35.16 | 45.19 | 13.34 | 29.19 |
| DiffTSR | 39.55 | 12.81 | 19.35 | 14.27 | 10.71 | 51.76 | 14.39 | 22.51 | 15.97 | 12.68 | 78.98 | 17.94 | 29.23 | 5.49 | 17.12 |
| InvSR | 55.56 | 60.22 | 57.79 | 27.53 | 40.64 | 65.44 | 57.05 | 60.96 | 31.65 | 43.84 | 78.67 | 55.38 | 65.00 | 15.95 | 48.15 |
| MARCONet | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 12.50 | 0.12 | 0.25 | 2.02 | 0.12 | 0.00 | 0.00 | 0.00 | 0.08 | 0.00 |
| MARCONet++ | 55.08 | 45.85 | 50.05 | 27.74 | 33.37 | 68.58 | 53.14 | 59.88 | 32.49 | 42.73 | 81.76 | 55.19 | 65.90 | 15.01 | 49.14 |
| PiSA-SR | 60.16 | **66.79** | **63.30** | 26.71 | **46.31** | 66.84 | **63.70** | 65.23 | 33.74 | 48.40 | 83.20 | 57.14 | 67.75 | 15.44 | 51.23 |
| Real-ESRGAN | 59.41 | 58.89 | 59.15 | **30.16** | 42.00 | **75.05** | 61.04 | **67.32** | 33.50 | **50.74** | 83.70 | 63.99 | 72.53 | 16.25 | 56.90 |
| SwinIR | 58.21 | 51.43 | 54.61 | 28.31 | 37.56 | 72.41 | 56.59 | 63.53 | **34.02** | 46.55 | 83.54 | 64.86 | 73.03 | 16.22 | 57.51 |
| StableSR | **62.08** | 57.83 | 59.88 | **30.32** | 42.73 | **73.79** | 56.13 | 63.76 | **34.71** | 46.80 | **84.70** | **65.56** | **73.91** | **16.81** | **58.62** |
| SUPIR | 58.16 | 58.67 | 58.41 | 20.17 | 41.26 | 64.54 | 59.48 | 61.90 | 26.81 | 44.83 | 74.54 | 53.28 | 62.14 | 15.86 | 45.07 |
| TAIR | 42.11 | 20.13 | 27.23 | 20.81 | 15.76 | 46.60 | 22.26 | 30.13 | 23.61 | 17.73 | 63.71 | 21.88 | 32.58 | 10.04 | 19.46 |
| GLYPH-SR (ours) | **61.33** | **75.14** | **67.54** | 22.17 | **50.99** | 68.07 | **75.79** | **71.72** | 28.37 | **55.91** | 79.22 | **68.07** | **73.22** | **19.49** | **57.76** |
| **SCUT-CTW1500(x4)** | | | | | | | | | | | | | | | |
| LR | 49.86 | 8.57 | 14.63 | 21.56 | 7.89 | 61.44 | 14.57 | 23.55 | 24.51 | 13.35 | 80.62 | 33.40 | 47.23 | 21.73 | 30.91 |
| HR | 72.57 | 68.73 | 70.59 | 56.36 | 54.55 | 75.74 | 65.33 | 70.15 | 52.17 | 54.02 | 87.67 | 69.89 | 77.77 | 35.27 | 63.63 |
| BSRGAN | 46.41 | 16.80 | 24.67 | 29.86 | 14.07 | 56.71 | 13.54 | 21.86 | 23.70 | 12.27 | 78.96 | 22.56 | 35.10 | 17.25 | 21.28 |
| DiffBIR | 38.18 | 18.26 | 24.71 | 33.85 | 14.09 | 36.43 | 17.70 | 23.82 | 30.71 | 13.52 | 54.93 | 21.31 | 30.71 | 20.94 | 18.14 |
| DiffTSR | 45.86 | 12.60 | 19.77 | 25.83 | 10.97 | 50.84 | 9.48 | 15.98 | 18.64 | 8.69 | 72.82 | 14.14 | 23.69 | 11.30 | 13.43 |
| InvSR | 45.37 | 21.93 | 29.57 | 34.39 | 17.35 | 47.40 | 18.31 | 26.41 | 28.17 | 15.22 | 66.15 | 23.33 | 34.50 | 18.34 | 20.84 |
| MARCONet | 17.65 | 0.07 | 0.13 | 0.71 | 0.07 | 13.83 | 0.29 | 0.57 | 2.55 | 0.29 | 41.67 | 0.11 | 0.22 | 0.20 | 0.11 |
| MARCONet++ | 47.02 | 14.98 | 22.72 | 27.05 | 12.82 | 54.82 | 12.70 | 20.63 | 22.17 | 11.50 | 74.92 | 21.24 | 33.10 | 15.93 | 19.83 |
| PiSA-SR | 49.11 | **30.27** | **37.46** | **40.32** | **23.04** | 56.25 | **24.50** | **34.14** | **33.47** | **20.58** | 71.18 | 31.96 | **44.11** | 23.23 | **28.30** |
| Real-ESRGAN | **52.95** | 22.22 | 31.31 | 33.69 | 18.56 | **59.50** | 17.41 | 26.94 | 26.49 | 15.57 | **79.94** | 29.65 | 43.25 | 20.12 | 27.59 |
| SwinIR | 49.75 | 15.05 | 23.10 | 28.11 | 13.06 | **59.40** | 14.42 | 23.21 | 24.64 | 13.13 | **80.24** | 26.14 | 39.44 | 17.58 | 24.56 |
| StableSR | **53.58** | 16.77 | 25.55 | 30.02 | 14.64 | 57.67 | 12.06 | 19.95 | 22.18 | 11.08 | 79.31 | **32.25** | **45.86** | 21.07 | **29.75** |
| SUPIR | 39.95 | 11.84 | 18.26 | 25.73 | 10.05 | 45.16 | 10.93 | 17.61 | 21.40 | 9.65 | 62.60 | 15.13 | 24.37 | 14.32 | 13.87 |
| TAIR | 46.43 | 26.80 | 33.98 | **38.09** | 20.47 | 50.58 | 20.76 | 29.44 | 32.51 | 17.26 | 66.76 | 30.29 | 41.67 | **23.32** | 26.32 |
| GLYPH-SR (ours) | 48.82 | **31.46** | **38.26** | 37.75 | **23.66** | 47.45 | **30.27** | **36.96** | **36.09** | **22.67** | 63.59 | **32.37** | 42.90 | **25.86** | 27.31 |
| **CUTE(x4)** | | | | | | | | | | | | | | | |
| LR | 72.41 | 60.29 | 65.80 | 51.93 | 49.03 | 75.66 | 52.27 | 50.58 | 50.32 | 44.75 | 93.72 | 79.51 | 80.47 | 38.43 | 75.49 |
| HR | 77.78 | 80.38 | 79.06 | 51.61 | 65.37 | 75.33 | 51.36 | 61.08 | 45.35 | 43.97 | 95.17 | 79.76 | 86.78 | 37.54 | 76.65 |
| BSRGAN | 68.84 | **77.89** | **73.09** | **54.63** | **57.59** | 69.44 | **46.95** | 56.02 | 45.37 | 38.91 | 92.54 | 76.86 | 83.97 | 39.00 | 72.37 |
| DiffBIR | 64.90 | 73.37 | 68.88 | 48.01 | 52.53 | 61.48 | 40.49 | 48.82 | 43.45 | 32.30 | 88.12 | 76.39 | 81.84 | 38.53 | 69.26 |
| DiffTSR | 64.94 | 57.65 | 61.08 | 51.95 | 43.97 | 67.80 | 36.53 | 47.48 | **45.54** | 31.13 | 92.59 | 61.22 | 73.71 | 30.66 | 58.37 |
| InvSR | 70.19 | 74.87 | 72.46 | 53.54 | 56.81 | **72.79** | 45.00 | 55.62 | 43.91 | 38.52 | 90.87 | **79.41** | **84.75** | 37.15 | **73.54** |
| MARCONet | 100.0 | 1.17 | 2.31 | 0.78 | 1.17 | 54.55 | 2.38 | 4.56 | 3.74 | 2.33 | **100.00** | 1.95 | 3.82 | 0.96 | 1.95 |
| MARCONet++ | 68.34 | 70.10 | 69.21 | 54.17 | 52.92 | 71.64 | 43.84 | 54.39 | 44.30 | 37.35 | 93.09 | 71.72 | 81.02 | **39.79** | 68.09 |
| PiSA-SR | 71.36 | 74.24 | 72.77 | 50.28 | 57.20 | 70.29 | 44.91 | 54.80 | 42.70 | 37.74 | **93.30** | 74.18 | 82.65 | 38.00 | 70.43 |
| Real-ESRGAN | **71.43** | 76.14 | **73.71** | 53.32 | **58.37** | 71.81 | **49.77** | **58.79** | 45.31 | **41.63** | 93.03 | 76.95 | 84.23 | 36.37 | 72.76 |
| SwinIR | **74.63** | 72.82 | **73.71** | **57.77** | **58.37** | 71.74 | 45.41 | 55.62 | **46.40** | 38.52 | 91.92 | 75.52 | 82.92 | 37.52 | 70.82 |
| StableSR | 69.71 | 74.74 | 72.14 | 51.76 | 56.42 | **74.64** | 46.40 | **57.22** | 42.36 | **40.08** | 89.66 | 77.12 | 82.92 | 38.02 | 70.82 |
| SUPIR | 68.78 | 73.06 | 70.85 | 49.43 | 54.86 | 63.38 | 43.90 | 51.87 | 42.38 | 35.02 | 89.05 | 76.17 | 82.11 | **40.24** | 69.65 |
| TAIR | 50.00 | 41.21 | 55.21 | 39.41 | 29.18 | 60.83 | 34.76 | 43.77 | 41.46 | 28.40 | 87.18 | 57.38 | 69.87 | 35.23 | 52.92 |
| GLYPH-SR (ours) | 69.48 | **77.08** | **73.09** | 47.00 | **57.59** | 68.28 | 46.92 | 55.62 | 38.27 | 38.52 | 90.05 | **80.51** | **85.01** | 39.78 | **73.93** |
| **SVT(x8)** | | | | | | | | | | | | | | | |
| LR | 34.62 | 4.84 | 8.49 | 8.92 | 4.43 | 53.04 | 18.82 | 27.78 | 19.86 | 16.13 | 79.78 | 29.23 | 42.79 | 8.65 | 27.22 |
| HR | 74.73 | 87.86 | 80.76 | 27.69 | 67.73 | 81.18 | 82.97 | 82.06 | 34.80 | 69.58 | 87.68 | 82.14 | 84.82 | 17.77 | 73.65 |
| BSRGAN | 35.75 | 9.18 | 14.61 | 13.02 | 7.88 | 36.54 | 7.99 | 13.12 | 14.68 | 7.02 | 76.77 | 15.34 | 25.56 | 6.25 | 14.66 |
| DiffBIR | 25.00 | 12.54 | 16.70 | 16.56 | 9.11 | 29.23 | 13.58 | 18.55 | 18.51 | 10.22 | 44.16 | 14.93 | 22.32 | 10.21 | 12.56 |
| DiffTSR | 28.03 | 6.29 | 10.28 | 11.36 | 5.42 | 31.51 | 6.46 | 10.72 | 14.98 | 5.67 | 62.50 | 9.09 | 15.87 | 4.60 | 8.62 |
| InvSR | 29.34 | 12.08 | 17.12 | **18.54** | 9.36 | 37.80 | 14.68 | 21.15 | 19.87 | 11.82 | 50.00 | 13.73 | 21.54 | 8.20 | 12.07 |
| MARCONet | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 2.30 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.00 |
| MARCONet++ | 28.03 | 6.29 | 10.28 | 12.41 | 5.42 | 38.36 | 8.54 | 13.97 | 14.26 | 7.51 | 62.96 | 15.60 | 22.32 | 6.62 | 14.29 |
| PiSA-SR | 36.11 | 11.57 | 17.53 | 14.53 | 9.61 | 47.84 | 16.06 | 24.05 | 19.83 | 13.67 | **79.41** | 24.77 | 37.76 | 7.72 | 23.28 |
| Real-ESRGAN | 34.50 | 11.93 | 17.73 | 16.17 | 9.73 | 48.20 | 15.35 | 23.29 | 19.45 | 13.18 | 76.68 | 19.30 | 30.83 | 7.14 | 18.23 |
| SwinIR | 34.59 | 9.26 | 14.61 | 13.61 | 7.88 | 45.63 | 13.43 | 20.75 | 17.89 | 11.58 | 76.44 | 19.04 | 30.48 | 6.45 | 17.98 |
| StableSR | 41.13 | 14.05 | 20.95 | 17.50 | 11.70 | **50.45** | 16.12 | 24.43 | 19.15 | 13.92 | **79.43** | **29.71** | **43.24** | 9.96 | **27.59** |
| SUPIR | **42.82** | **27.66** | **33.61** | 15.29 | **20.20** | 43.00 | **30.90** | **35.96** | 18.80 | **21.92** | 59.22 | 26.68 | 36.78 | **11.28** | 22.54 |
| TAIR | 35.38 | 15.48 | 21.54 | 16.68 | 12.07 | 40.91 | 16.46 | 23.48 | **21.45** | 13.30 | 51.87 | 17.96 | 26.68 | 10.14 | 15.39 |
| GLYPH-SR (ours) | **48.52** | **49.06** | **48.79** | **19.16** | **32.27** | **57.32** | **55.03** | **56.16** | **23.17** | **39.04** | 69.57 | **50.53** | **58.54** | **17.99** | **41.38** |
| **SCUT-CTW1500(x8)** | | | | | | | | | | | | | | | |
| LR | 27.91 | 0.27 | 0.53 | 1.38 | 0.26 | 38.41 | 2.54 | 4.76 | 8.22 | 2.44 | 73.05 | 5.47 | 10.18 | 5.14 | 5.36 |
| HR | 72.57 | 68.73 | 70.59 | 56.36 | 54.55 | 75.74 | 65.33 | 70.15 | 52.17 | 54.02 | 87.67 | 69.89 | 77.77 | 35.27 | 63.63 |
| BSRGAN | 29.10 | 1.79 | 3.37 | 7.67 | 1.72 | 31.06 | 1.88 | 3.54 | 7.31 | 1.80 | 64.75 | 2.00 | 3.88 | 2.10 | 1.98 |
| DiffBIR | 15.66 | 2.81 | 4.76 | 17.18 | 2.44 | 11.46 | 3.28 | 5.10 | 16.05 | 2.62 | 21.26 | 2.60 | 4.64 | 9.28 | 2.37 |
| DiffTSR | 28.10 | 1.55 | 2.95 | 6.94 | 1.50 | 28.33 | 1.51 | 2.86 | 6.44 | 1.45 | 51.94 | 1.49 | 2.90 | 2.37 | 1.47 |
| InvSR | 21.15 | 1.10 | 2.09 | 7.13 | 1.06 | 18.66 | 1.15 | 2.17 | 7.17 | 1.10 | 55.45 | 1.24 | 2.43 | 1.78 | 1.23 |
| MARCONet | 20.00 | 0.07 | 0.13 | 0.59 | 0.07 | 16.67 | 0.31 | 0.61 | 2.28 | 0.31 | 46.15 | 0.13 | 0.26 | 0.33 | 0.13 |
| MARCONet++ | 35.53 | 1.21 | 2.35 | 4.27 | 1.19 | 34.48 | 1.35 | 2.60 | 5.15 | 1.32 | 67.42 | 1.33 | 2.60 | 1.03 | 1.32 |
| PiSA-SR | 25.50 | 4.48 | 7.61 | 17.41 | 3.96 | 29.21 | 3.92 | 6.92 | 13.26 | 3.58 | 50.45 | 5.20 | 9.43 | 6.82 | 4.95 |
| Real-ESRGAN | 32.77 | 2.72 | 5.02 | 9.82 | 2.57 | **34.74** | 3.07 | 5.64 | 9.91 | 2.90 | 66.30 | 4.11 | 7.74 | 3.81 | 4.02 |
| SwinIR | **36.32** | 1.93 | 3.67 | 6.60 | 1.87 | 32.06 | 2.52 | 4.68 | 9.01 | 2.40 | **67.96** | 2.74 | 5.27 | 2.77 | 2.70 |
| StableSR | **39.90** | 1.74 | 3.33 | 5.63 | 1.69 | **40.55** | 2.34 | 4.43 | 7.41 | 2.26 | **71.37** | 3.95 | 7.49 | 3.68 | 3.89 |
| SUPIR | 19.91 | 3.15 | 5.43 | 14.12 | 2.79 | 22.41 | 3.64 | 6.26 | 13.13 | 3.23 | 33.47 | 3.91 | 7.00 | 7.13 | 3.63 |
| TAIR | 27.45 | **6.67** | **10.74** | **21.66** | **5.67** | 28.06 | **5.52** | **9.23** | **18.45** | **4.84** | 38.99 | **7.19** | **12.14** | **12.97** | **6.46** |
| GLYPH-SR (ours) | 22.61 | **7.35** | **11.09** | **20.54** | **5.87** | 25.24 | **10.38** | **14.71** | **20.10** | **7.94** | 34.12 | **9.34** | **14.67** | **13.85** | **7.92** |
| **CUTE(x8)** | | | | | | | | | | | | | | | |
| LR | 60.00 | 29.30 | 39.38 | 38.33 | 24.51 | 68.33 | 37.44 | 36.31 | 45.45 | 31.91 | 91.02 | 62.81 | 67.18 | 29.99 | 59.14 |
| HR | 77.78 | 80.38 | 79.06 | 51.61 | 65.37 | 75.33 | 51.36 | 61.08 | 45.35 | 43.97 | 95.17 | 79.76 | 86.78 | 37.54 | 76.65 |
| BSRGAN | 58.33 | 52.41 | 55.21 | 47.56 | 38.13 | **68.42** | 35.29 | **46.57** | **42.81** | **30.35** | 91.61 | 58.20 | 71.18 | 28.40 | 55.25 |
| DiffBIR | 61.58 | 57.67 | **59.56** | 45.10 | **42.41** | 58.73 | 36.10 | 44.71 | 38.97 | 28.79 | 88.61 | 58.58 | 70.53 | 30.10 | 54.47 |
| DiffTSR | 60.76 | 49.23 | 54.39 | 48.20 | 37.35 | 62.16 | 32.09 | 42.33 | 41.04 | 26.85 | 86.23 | 50.00 | 63.30 | 27.95 | 46.30 |
| InvSR | 55.80 | 57.06 | 56.42 | 47.28 | 39.30 | 60.98 | 35.89 | 45.18 | 39.50 | 29.18 | 87.43 | 61.86 | 72.46 | **35.41** | 56.81 |
| MARCONet | **100.00** | 1.56 | 3.07 | 1.15 | 1.56 | 33.33 | 2.45 | 4.56 | 4.65 | 2.33 | **100.00** | 1.95 | 3.82 | 0.83 | 1.95 |
| MARCONet++ | 56.49 | 45.79 | 50.58 | 48.26 | 33.85 | 67.57 | 33.94 | 45.18 | 40.53 | 29.18 | 90.51 | 59.09 | 71.50 | 33.08 | 55.64 |
| PiSA-SR | 58.23 | 48.17 | 52.72 | **51.39** | 35.80 | 61.61 | 32.24 | 42.33 | 41.76 | 26.85 | 92.26 | **63.52** | **75.24** | 30.23 | **60.31** |
| Real-ESRGAN | 60.67 | 57.75 | 59.18 | 50.53 | 42.02 | **70.00** | **38.01** | **49.27** | **42.81** | **32.68** | **93.25** | 61.79 | **74.33** | 31.11 | **59.14** |
| SwinIR | 58.33 | 47.40 | 52.30 | 49.71 | 35.41 | 64.66 | 34.72 | 45.18 | 42.14 | 29.18 | 90.68 | 60.33 | 72.46 | 28.55 | 56.81 |
| StableSR | 60.06 | 55.73 | 57.81 | **51.68** | 40.66 | 61.22 | 35.80 | 45.18 | **43.35** | 29.18 | 88.79 | 63.24 | 73.87 | 32.52 | 58.56 |
| SUPIR | 57.69 | **58.33** | 58.01 | 43.46 | 40.86 | 59.83 | 33.33 | 42.81 | 35.98 | 27.24 | 82.25 | 61.23 | 70.20 | **35.16** | 54.09 |
| TAIR | 49.30 | 37.84 | 42.81 | 37.98 | 27.24 | 54.10 | 32.84 | 40.87 | 40.20 | 25.68 | 82.27 | 50.00 | 62.20 | 33.88 | 45.14 |
| GLYPH-SR (ours) | **63.49** | **63.83** | **63.66** | 42.40 | **46.69** | 58.91 | **37.25** | 45.65 | 36.80 | 29.57 | 83.80 | **65.79** | 73.71 | 35.12 | 58.37 |

Table 3: Quantitative comparison of SR models on three scene text datasets (SVT, SCUT-CTW1500, CUTE80) at ×4 and ×8 upscaling factors. Metrics include distortion-based (PSNR, SSIM, LPIPS↓) and perceptual quality scores (MANIQA, CLIP-IQA, MUSIQ). Red and blue denote the best and second-best results, respectively.

| Dataset | SR model | PSNR | SSIM | LPIPS↓ | MANIQA | CLIP-IQA | MUSIQ |
|---|---|---|---|---|---|---|---|
| SVT(x4) | LR | 30.12 | 87.21 | 33.58 | 20.45 | 17.06 | 26.31 |
| | HR | - | - | - | 23.64 | 25.51 | 41.63 |
| | BSRGAN | 28.09 | 83.16 | 35.34 | 38.16 | 39.63 | 66.25 |
| | DiffBIR | 21.96 | 63.94 | 43.55 | **47.82** | **58.66** | **71.18** |
| | DiffTSR | 26.06 | 78.42 | 44.95 | 21.34 | 27.69 | 46.24 |
| | InvSR | 24.78 | 76.58 | 38.61 | 46.78 | 57.30 | 70.81 |
| | PiSA-SR | 26.58 | 82.04 | 34.13 | 37.41 | 44.30 | 61.87 |
| | MARCONet | 18.40 | 69.07 | 58.91 | 30.92 | 24.43 | 27.26 |
| | MARCONet++ | 27.73 | 84.33 | 34.75 | 29.31 | 19.82 | 49.20 |
| | Real-ESRGAN | 29.67 | **88.58** | **30.68** | 31.16 | 28.58 | 51.14 |
| | SwinIR | **30.48** | 86.38 | 35.29 | 26.32 | 44.50 | 34.55 |
| | StableSR | **30.54** | **87.00** | **33.73** | 24.75 | 32.18 | 24.44 |
| | SUPIR | 22.76 | 67.15 | 45.14 | 42.36 | 48.42 | 67.55 |
| | TAIR | 24.80 | 73.72 | 40.31 | 29.31 | 19.82 | 49.20 |
| | GLYPH-SR (ours) | 22.89 | 67.19 | 42.20 | **47.75** | **59.40** | **70.99** |
| SCUT-CTW1500(x4) | LR | 19.16 | 55.44 | 47.91 | 28.92 | 31.16 | 25.82 |
| | HR | - | - | - | 64.23 | 77.42 | 70.87 |
| | BSRGAN | **20.22** | **64.59** | 32.12 | 51.41 | 47.44 | 67.52 |
| | DiffBIR | 17.91 | 56.34 | 36.20 | 62.37 | **61.90** | **71.19** |
| | DiffTSR | 18.99 | 58.59 | 41.34 | 35.39 | 30.59 | 55.83 |
| | InvSR | 18.32 | 60.71 | **32.99** | 57.75 | 55.94 | 69.25 |
| | PiSA-SR | 20.07 | 63.99 | **31.18** | 56.31 | 53.05 | 68.19 |
| | MARCONet | 14.80 | 40.89 | 66.30 | 33.34 | 16.54 | 28.78 |
| | MARCONet++ | 19.08 | 58.77 | 41.60 | 34.65 | 19.58 | 43.61 |
| | Real-ESRGAN | **20.85** | **67.46** | 36.81 | 40.81 | 43.43 | 52.66 |
| | SwinIR | 19.91 | 58.45 | 47.11 | 33.85 | 46.07 | 39.36 |
| | StableSR | 19.24 | 55.45 | 49.03 | 31.04 | 43.61 | 24.92 |
| | SUPIR | 13.61 | 32.98 | 52.15 | 57.35 | 51.68 | 66.96 |
| | TAIR | 18.19 | 60.04 | 34.46 | **65.38** | 47.05 | 67.08 |
| | GLYPH-SR (ours) | 18.19 | 54.67 | 37.15 | **70.33** | **57.88** | **70.31** |
| CUTE80(x4) | LR | 26.44 | 78.35 | 36.69 | 28.93 | 36.80 | 37.64 |
| | HR | - | - | - | 40.12 | 55.71 | 60.81 |
| | BSRGAN | **27.35** | **79.76** | 31.83 | 44.22 | 55.73 | 69.13 |
| | DiffBIR | 22.60 | 66.07 | 37.74 | **51.04** | **72.64** | 69.06 |
| | DiffTSR | 24.06 | 72.66 | 42.74 | 33.94 | 38.47 | 58.74 |
| | InvSR | 24.41 | 75.55 | 32.93 | 50.30 | **67.78** | **70.66** |
| | PiSA-SR | 25.83 | 77.41 | **31.49** | 45.82 | 61.81 | 66.18 |
| | MARCONet | 16.17 | 63.52 | 56.18 | 33.58 | 26.69 | 31.06 |
| | MARCONet++ | 25.16 | 77.75 | 34.91 | 31.88 | 34.90 | 54.15 |
| | Real-ESRGAN | **28.14** | **82.30** | 32.01 | 38.20 | 48.71 | 60.65 |
| | SwinIR | 27.18 | 77.95 | 38.02 | 31.87 | 59.32 | 47.94 |
| | StableSR | 26.23 | 79.51 | **30.45** | 36.26 | 49.74 | 60.09 |
| | SUPIR | 22.42 | 66.20 | 39.33 | 47.50 | 62.62 | 68.26 |
| | TAIR | 20.82 | 69.03 | 41.27 | **58.25** | 49.76 | **72.06** |
| | GLYPH-SR (ours) | 23.03 | 69.54 | 37.03 | 49.77 | 65.93 | 69.96 |
| SVT(x8) | LR | 26.15 | 78.90 | 48.54 | 19.81 | 44.07 | 22.96 |
| | HR | - | - | - | 23.64 | 25.51 | 41.63 |
| | BSRGAN | 25.13 | 73.71 | 45.64 | 37.14 | 37.58 | 62.83 |
| | DiffBIR | 22.89 | 65.20 | 50.07 | **45.54** | **53.20** | 64.11 |
| | DiffTSR | 24.45 | 76.19 | 46.32 | 21.39 | 26.39 | 43.96 |
| | InvSR | 22.82 | 71.34 | **41.84** | 32.51 | 50.83 | 51.69 |
| | PiSA-SR | 26.12 | 77.64 | 50.83 | 34.02 | 18.39 | 30.24 |
| | MARCONet | 18.68 | 69.49 | 58.71 | 30.84 | 24.76 | 27.02 |
| | MARCONet++ | 25.15 | 76.11 | 44.44 | 20.97 | 8.44 | 38.06 |
| | Real-ESRGAN | 25.69 | **80.28** | **41.92** | 28.38 | 17.86 | 43.01 |
| | SwinIR | **26.48** | 78.16 | 48.01 | 22.05 | 26.68 | 30.33 |
| | StableSR | **26.38** | **78.15** | 50.20 | 23.16 | 23.38 | 16.22 |
| | SUPIR | 21.23 | 59.08 | 51.46 | 40.17 | 45.06 | **65.20** |
| | TAIR | 22.72 | 68.24 | 44.86 | 31.99 | 29.49 | 54.27 |
| | GLYPH-SR (ours) | 21.77 | 61.36 | 47.85 | **47.40** | **56.78** | **69.93** |
| SCUT-CTW1500(x8) | LR | 17.04 | 43.70 | 63.73 | 16.39 | 26.19 | 17.71 |
| | HR | - | - | - | 64.23 | 77.42 | 70.87 |
| | BSRGAN | 17.32 | **48.50** | 47.86 | 46.21 | 37.83 | **66.05** |
| | DiffBIR | 15.78 | 43.47 | **50.05** | 54.75 | **49.89** | 63.16 |
| | DiffTSR | 14.83 | 40.25 | 54.50 | 35.49 | 31.88 | 50.43 |
| | InvSR | 11.81 | 30.68 | 65.88 | 29.65 | 29.62 | 40.29 |
| | PiSA-SR | 17.22 | 47.63 | **48.90** | 41.77 | 36.75 | 58.95 |
| | MARCONet | 14.93 | 40.72 | 66.71 | 33.56 | 16.20 | 28.95 |
| | MARCONet++ | 16.79 | 44.04 | 57.88 | 14.75 | 8.06 | 30.27 |
| | Real-ESRGAN | **17.65** | **52.34** | 52.14 | 28.37 | 20.95 | 39.99 |
| | SwinIR | **17.39** | 45.42 | 61.26 | 19.00 | 24.14 | 25.64 |
| | StableSR | 17.00 | 43.50 | 66.02 | 20.93 | 20.92 | 16.62 |
| | SUPIR | 12.63 | 26.51 | 58.63 | 55.46 | 47.02 | 65.55 |
| | TAIR | 16.02 | 45.70 | 45.73 | **63.60** | 36.57 | **66.38** |
| | GLYPH-SR (ours) | 16.27 | 41.31 | 52.58 | **61.94** | **48.21** | 63.43 |
| CUTE80(x8) | LR | 23.25 | 71.52 | 47.61 | 17.29 | 22.58 | 17.32 |
| | HR | - | - | - | 40.12 | 55.71 | 60.81 |
| | BSRGAN | **23.84** | **72.55** | **39.14** | 42.07 | 54.31 | **67.33** |
| | DiffBIR | 22.77 | 65.36 | 41.79 | **47.53** | 62.09 | 64.62 |
| | DiffTSR | 22.67 | 70.41 | 42.80 | 33.55 | 42.95 | 57.47 |
| | InvSR | 21.83 | 70.76 | **38.05** | 37.66 | **62.43** | 57.69 |
| | PiSA-SR | 23.36 | 70.52 | 47.71 | 30.71 | 30.80 | 45.16 |
| | MARCONet | 16.37 | 63.64 | 56.08 | 33.66 | 26.69 | 30.88 |
| | MARCONet++ | 22.79 | 70.85 | 43.34 | 21.03 | 22.72 | 44.24 |
| | Real-ESRGAN | **24.01** | **75.58** | 39.60 | 35.17 | 36.46 | 56.55 |
| | SwinIR | 23.82 | 71.22 | 47.71 | 22.72 | 40.40 | 39.44 |
| | StableSR | 7.94 | 35.66 | 79.75 | 26.00 | 40.42 | 34.48 |
| | SUPIR | 20.64 | 61.31 | 43.76 | 46.38 | 61.67 | 67.04 |
| | TAIR | 20.49 | 67.69 | 42.35 | 37.11 | 36.84 | 55.06 |
| | GLYPH-SR (ours) | 21.19 | 65.15 | 42.31 | **47.75** | **65.85** | **68.85** |

Table 4: Ablation on the scheduler policy evaluated on the CUTE80 dataset.

(a) CUTE80 (LR × 4)

| Scheduler Policy | MANIQA | CLIP-IQA | MUSIQ | OCR $F_1$ |
|---|---|---|---|---|
| Binary ping–pong | 49.77 | 65.93 | 69.96 | **85.01** |
| Mixing ($\lambda_t = 0.1$) | **49.95** | **70.64** | **70.67** | 81.57 |
| Mixing ($\lambda_t = 0.3$) | 49.04 | 69.56 | 69.75 | 83.18 |
| Mixing ($\lambda_t = 0.5$) | 47.57 | 65.47 | 68.95 | 84.23 |
| Mixing ($\lambda_t = 0.7$) | 47.86 | 68.91 | 68.83 | 81.84 |
| Mixing ($\lambda_t = 0.9$) | 48.85 | 69.11 | 69.13 | 82.65 |

(b) CUTE80 (LR × 8)

| Scheduler Policy | MANIQA | CLIP-IQA | MUSIQ | OCR $F_1$ |
|---|---|---|---|---|
| Binary ping–pong | 47.75 | 65.85 | 68.85 | **73.71** |
| Mixing ($\lambda_t = 0.1$) | **48.89** | 67.65 | **69.56** | 66.49 |
| Mixing ($\lambda_t = 0.3$) | 47.44 | **68.31** | 68.86 | 69.87 |
| Mixing ($\lambda_t = 0.5$) | 46.57 | 64.07 | 67.35 | 73.40 |
| Mixing ($\lambda_t = 0.7$) | 45.80 | 67.98 | 67.19 | 66.84 |
| Mixing ($\lambda_t = 0.9$) | 45.58 | 67.66 | 67.18 | 68.88 |

Table 5: Ablation study on CFG scale $w$. Evaluated on CUTE80 (×4).

| CFG Scale ($w$) | OCR Metric ($F_1$ Score ↑) | | | SR Metric (↑) | | |
|---|---|---|---|---|---|---|
| | OpenOCR | GOT-OCR | LLaVA-NeXT | MANIQA | CLIP-IQA | MUSIQ |
| 2 | **75.54** | **57.62** | 87.53 | 41.26 | 55.96 | 65.23 |
| 4 | 74.63 | 57.22 | **88.50** | 43.42 | 60.01 | 67.30 |
| **7.5*** | 73.09 | 55.62 | 85.01 | 49.77 | 65.93 | 69.96 |
| 12 | 71.50 | 53.56 | 84.23 | 49.52 | 69.87 | 70.51 |
| 20 | 65.45 | 45.18 | 79.06 | **50.60** | **74.30** | **71.03** |

Table 6: Ablation study on Control Scale $s_{\text{CTRL}}$. Evaluated on CUTE80 (×4).

| $s_{CTRL}$ | OCR Metric ($F_1$ Score ↑) | | | SR Metric (↑) | | |
|---|---|---|---|---|---|---|
| | OpenOCR | GOT-OCR | LLaVA-NeXT | MANIQA | CLIP-IQA | MUSIQ |
| **1.0*** | **73.09** | **55.62** | **85.01** | **49.77** | 65.93 | **69.96** |
| 2.0 | 69.21 | 56.02 | 81.84 | 26.36 | **66.95** | 49.92 |
| 3.0 | 4.56 | 6.04 | 25.17 | 37.17 | 45.28 | 49.21 |
| 10.0 | 0.00 | 0.00 | 0.00 | 35.22 | 25.26 | 27.18 |

Table 7: **Compute comparison.** For VLM-guided methods, #Params lists (*restoration*, *VLM*) in millions.

| Method | #Params (M) | Inference (s / sample) | Peak VRAM (GB) |
|---|---|---|---|
| StableSR | 153 | $79.98 \pm 0.22$ | 10.10 |
| DiffBIR | 385 | $53.14 \pm 1.41$ | 9.64 |
| SUPIR | 18, 152 | $25.25 \pm 0.86$ | 46.21 |
| GLYPH-SR | 13, 225 | $38.25 \pm 1.28$ | 43.56 |

| Dataset / Guider | OCR metric: $F_1$ ↑ | | | SR metric | | |
|---|---|---|---|---|---|---|
| | OpenOCR | GOT-OCR | LLaVA-NeXT | MANIQA | CLIP-IQA | MUSIQ |
| Ours | 73.09 | 55.62 | 85.01 | 49.77 | 65.93 | 69.96 |
| R-4B | 72.77 | 55.21 | 84.75 | 40.78 | 57.20 | 65.57 |
| PaliGemma-3B | 70.53 | 51.45 | 84.52 | 42.11 | 59.61 | 65.83 |

Table 8: Compact/alternative guiders in GLYPH-SR.

Table 9: Preliminary evaluation on CUTE80 (×16) comparing the baseline GLYPH-SR with proposed solution strategies. **Red** indicates improvement, and **Blue** indicates degradation compared to the baseline.

| **Method** | **OpenOCR** | **GOT-OCR** | **ManiQA** | **ClipIQA** |
|---|---|---|---|---|
| LR | 12.41 | 12.41 | 14.12 | 29.11 |
| GLYPH-SR (Baseline) | 9.63 | 11.72 | 38.52 | 54.68 |
| GLYPH-SR + Sung et al. (2025) | **13.77** (+4.14) | 12.41 (+0.69) | 35.75 (-2.77) | 49.68 (-5.00) |
| GLYPH-SR + CoT | 10.33 (+0.70) | **13.09** (+1.37) | **42.15** (+3.63) | **59.99** (+5.31) |

Table 10: **Trainable parameter counts** (millions). Despite using a VLM, GLYPH-SR keeps *trainable* parameters modest via freezing and LoRA.

| **Metric** | **GLYPH-SR** | **PiSA-SR** | **StableSR** | **DiffBIR** | **SUPIR** | **DiffTSR** |
|---|---|---|---|---|---|---|
| Trainable (M) | **60.7** | 0.38 | 152.67 | 378.95 | 3865.64 | 55.31 |