GLYPH-SR: CAN WE ACHIEVE BOTH HIGH-QUALITY IMAGE SUPER-RESOLUTION AND HIGH-FIDELITY TEXT RECOVERY VIA VLM-GUIDED LATENT DIFFUSION MODEL?

Anonymous authors

000

001

002

004

006

008

009

010 011 012

013

015

016

018

019

021

024

025

026

027

028

029

031

034

037

039 040

041

043

044

046

047

048

050 051

052

Paper under double-blind review

ABSTRACT

Image super-resolution (SR) is fundamental to 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 the most actionable information; when characters are blurred or hallucinated, optical character recognition (OCR) and subsequent decisions fail even if the rest of the image appears sharp. Yet previous SR research has often been tuned to distortion (PSNR/SSIM) or learned perceptual metrics (LPIPS, MANIQA, CLIP-IQA, MUSIQ) that are largely insensitive to character-level errors. Furthermore, studies that do address text SR often focus on simplified benchmarks with isolated characters, overlooking the challenges of text within complex natural scenes. 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 utilizes a Text-SR Fusion ControlNet (TS-ControlNet) guided by OCR data, and a ping-pong scheduler that alternates between text- and scene-centric guidance. To enable targeted text restoration, we train these components 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_I 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 satisfy both objectives simultaneously—high readability and high visual realism—delivering SR that looks right and reads right. We provide code, pretrained models, the synthetic corpus with generation scripts, and an evaluation suite to support reproducibility.

1 Introduction

Image super-resolution (SR), which reconstructs high-resolution (HR) images from low-resolution (LR) inputs, is critical for applications like autonomous driving where clear details are paramount. While conventional SR aims to improve perceptual quality, we argue that for many real-world scenarios, ensuring the text legibility of scene-text (e.g., on signs, license plates) is equally, if not more, 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 AN OVERLOOKED CHALLENGE IN IMAGE SR: ACHIEVING HIGH SCENE-TEXT FIDELITY

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.,

¹Throughout this paper, we will use *image SR* and *SR* interchangably whenever there is no ambiguity.

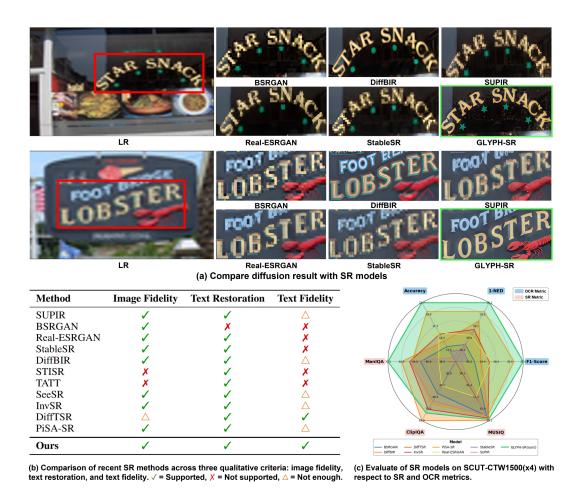


Figure 1: Qualitative and quantitative comparisons of our GLYPH-SR with other competing SR methods, demonstrating superior text fidelity and OCR F_I score.

StableSR Wang et al. (2024), DiffBIR Lin et al. (2024), InvSR Yue et al. (2025)) despite strong perceptual scores:

- (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; small text regions (often well below 1% of the image) therefore contribute little, so character corruption is 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 alongside mediocre perceptual quality. 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 that achieves both. Our key technical contributions and breakthroughs in this work include the followings:

- **Bi-Objective Formulation & Dual-Axis Evaluation.** We explicitly cast SR in text-rich scenes as the joint optimization of *image quality* and *readability*, and standardize a *dual-axis* protocol that reports perceptual SR metrics (MANIQA, CLIP-IQA, MUSIQ) *together with* OCR-aware measures (word/character accuracy, edit distance, F₁), 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}_{TXT} and a scene caption \mathcal{S}_{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 λ_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 ×4/×8, GLYPH-SR improves OCR F₁ by up to +15.18 pp 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 WORKS

SR via Deep Learning. Early CNN methods such as SRCNN Dong et al. (2015), EDSR Lim et al. (2017), and RCAN Zhang et al. (2018b), and later transformer models like SwinIR Liang et al. (2021), substantially advanced distortion-oriented SR; yet they primarily optimize pixel fidelity rather than semantic fidelity in small, text-bearing regions. Adversarially trained SR has improved perceptual realism on in-the-wild images; representative examples include BSRGAN Zhang et al. (2021) and Real-ESRGAN Wang et al. (2021).

Diffusion-based SR has recently shown strong stability and realism. Foundational approaches such as DiffBIR Lin et al. (2024) and StableSR Wang et al. (2024) couple LR conditioning with powerful diffusion priors, and subsequent work incorporates richer priors or auxiliary conditions: SeeSR Wu et al. (2024) exploits semantic prompts, InvSR Yue et al. (2025) enables flexible guidance/sampling, SUPIR Yu et al. (2024b) leverages large-scale pretrained backbones with restoration-guided sampling, and PISA-SR Sun et al. (2025) further advances controllability. As illustrated in Fig. 1(b), explicit character-level integrity is seldom a primary optimization target in general-purpose diffusion SR. Consequently, as further substantiated by the quantitative benchmarks in Fig. 1(c), there is a notable scarcity of methods that holistically address both general image fidelity and text-specific restoration metrics.

Text-Focused SR. Text-centric SR aims to enhance readability with text-aware priors or recognition-aware objectives. Representative methods include TATT Ma et al. (2022), STISR Noguchi et al. (2024), and Stroke-Aware SR Chen et al. (2022). While effective on word/line crops, these approaches often assume simplified settings and can underperform on full natural scenes where text must be preserved together with surrounding content.

3 Our Approach: GLYPH-SR

3.1 Model architecture

Overview. Fig. 2 depicts the proposed **GLYPH-SR** pipeline. Given an LR image $\mathbf{I}_{LR} \in \mathbb{R}^{H \times W \times C}$, an LR-robust conditioner of a pretrained latent diffusion model (LDM) Rombach et al. (2022) extracts multi-scale features f_{LR} used for conditioning. Our **Text–SR Fusion ControlNet** (**TS-ControlNet**) then injects complementary restoration cues while preserving the generative prior of the LDM. Finally, an Elucidated Diffusion Model (EDM) sampler Karras et al. (2022) drives the reverse process in latent space toward a high-resolution reconstruction. However, when guidance is provided only in a

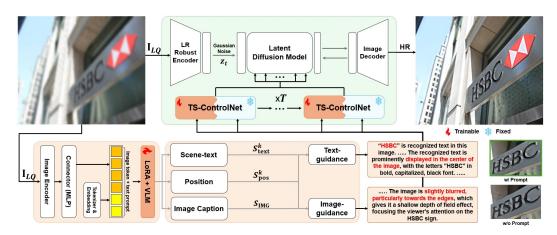


Figure 2: Overview of the proposed GLYPH-SR architecture.

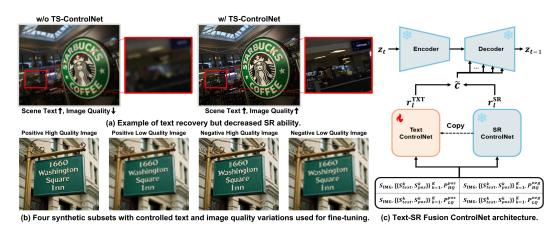


Figure 3: Text-centric fine-tuning framework: (a) trade-off between scene-text fidelity and overall image quality according to guidance; (b) four synthetic training subsets with matched prompts; (c) TS-ControlNet architecture.

holistic form, small text regions may still be treated as generic high-frequency textures rather than semantically meaningful glyphs, which can yield imperfect character restoration.

Condition Decomposition. To address this limitation, we explicitly separate the guidance into (i) **image-oriented** and (ii) **text-oriented** signals.

- Image-Oriented Guidance. A scene-level caption $S_{\rm IMG}$ summarizes global attributes such as illumination, composition, and depth-of-field, and is used to encourage holistic perceptual quality.
- Text-Oriented Guidance. A dedicated OCR module detects K text instances and returns position—text pairs $\{(\mathcal{S}^k_{\text{text}},\mathcal{S}^k_{\text{pos}})\}_{k=1}^K$. 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(b), simply separating $S_{\rm IMG}$ and $\{(S_{\rm text}^k, S_{\rm pos}^k)\}_{k=1}^K$ improves text fidelity but can degrade non-text regions, motivating our subsequent guidance-fusion strategy and the ping–pong scheduler that alternates text-centric and scene-centric guidance.

Text–SR Fusion ControlNet. To balance the two objectives—image quality and text legibility—we introduce the *Text–SR Fusion ControlNet* (TS-ControlNet), which merges glyph-level semantic priors with global SR guidance (Fig. 3c). During training, the LDM backbone and the SR branch of TS-

ControlNet are frozen, and only the text branch is updated, improving text legibility while preserving overall image quality.

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

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

The diffusion model \mathcal{D}_{θ} predicts the noise residual conditioned on two control streams: (i) \mathcal{C}_{SR} , a spatial condition from a frozen SR-ControlNet that guides the overall structure based on the low-resolution input image \mathcal{S}_{IMG} , and (ii) \mathcal{C}_{TXT} , a textual condition from a trainable Text-ControlNet that controls the rendering of text based on a set of OCR-derived text-position pairs \mathcal{S}_{TXT} .

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

Diffusion Loss with Residual Injection. The frozen SR-ControlNet and the trainable Text-ControlNet produce residual hierarchies. We blend them before injection via

$$c = \frac{1}{2} s_{\text{CTRL}} \Big[\mathcal{C}_{\text{SR}} (z_t; \phi_{\text{img}}(\mathcal{S}_{\text{IMG}} + P)) + \mathcal{C}_{\text{TXT}} (z_t; \phi_{\text{txt}}(\mathcal{S}_{\text{TXT}} + P)) \Big]. \tag{1}$$

where s_{CTRL} is a global scaling factor and P denotes the restoration guide prompt.

The diffusion backbone \mathcal{D}_{θ} then predicts the residual noise, and we optimize TS-ControlNet with the standard ε -prediction objective:

$$\mathcal{L}_{\text{text}} = \mathbb{E}_{z_0, t, \varepsilon} \| \varepsilon - \mathcal{D}_{\theta}(z_t, t, c) \|_2^2.$$
 (2)

Synthetic Fine-Tuning Dataset. To disentangle text legibility from holistic perceptual quality, we synthesize four mutually exclusive subsets $\left\{\mathbf{I}_{HQ}^{\mathrm{pos}},\mathbf{I}_{LQ}^{\mathrm{neg}},\mathbf{I}_{HQ}^{\mathrm{neg}},\mathbf{I}_{LQ}^{\mathrm{neg}}\right\}$. All synthetic data are generated from the same raw text, but for training purposes, the image quality is intentionally reduced or only the text within the images is distorted. As shown in Fig. 3 (b). To train *TS-ControlNet*, we defined the following guide prompt.

- **Positive–Text / High-Quality** ($P_{\rm HQ}^{\rm pos}$). Perfect image quality with perfectly preserved character outlines and precise positioning.
- Negative–Text / High-Quality ($P_{\rm HQ}^{\rm neg}$). Intentionally damaged character outlines and precise positioning, but good image quality.
- Positive–Text / Low-Quality (P_{LQ}^{pos}). Poor image quality, but preserved character outlines and precise positioning.
- Negative–Text / Low-Quality (P_{LQ}^{neg}). Image quality is poor and character outlines and exact positions are intentionally damaged.

Each sample is encoded into a *composite conditioning tuple* for the TS-ControlNet:

$$\underbrace{z_{\star}^{\diamond}}_{\text{image latent}} \oplus \underbrace{\psi(\mathcal{S}_{\text{IMG}})}_{\text{scene caption}} \oplus \underbrace{\psi(\{(\mathcal{S}_{\text{text}}^{k}, \mathcal{S}_{\text{pos}}^{k})\}_{k=1}^{K})}_{\text{text cues}} \oplus \underbrace{P_{\star}^{\diamond}}_{\text{guide prompt}}, \quad \diamond \in \{\text{pos, neg}\}, \; \star \in \{\text{HQ, LQ}\}.$$

Here, $z_{\star}^{\diamond} = \operatorname{Enc}(\mathbf{I}_{\star}^{\diamond})$ is the first-stage latent of the synthetic image $\mathbf{I}_{\star}^{\diamond}$, and $\psi(\cdot)$ denotes the frozen CLIP text encoder. Note that, to explicitly inform the model when incorrect text has been generated, the text-position pairs $\{(\mathcal{S}_{\text{text}}^k, \mathcal{S}_{\text{pos}}^k)\}_{k=1}^K$ are always extracted from the positive-text, high-quality image dataset.

3.2 Text-Image Balancing Scheduler

Although the dedicated *TS-ControlNet* injects glyph-centric features, the temporal allocation between text and image guidance along the diffusion trajectory is critical. We therefore introduce a scheduler

 $\mathcal{T}_{sched}: \{0,\ldots,T\} \to [0,1]$ that dynamically reweights the two guidance streams via a time-dependent coefficient λ_t .

paragraphStep update with mixed guidance. Let z_t be the latent at diffusion step t (sampling proceeds from t=T down to 0). Given a mixed embedding e^t (Eq. 4), we form a classifier-free guided noise estimate (Eq. 5) and then update

$$z_{t-1} = z_t - \eta_t \,\widehat{\epsilon}_t, \tag{3}$$

where η_t is a step size (a function of the noise level σ_t in our EDM-based solver). At inference we initialize $z_T \sim \mathcal{N}(0, \sigma_T^2 \mathbf{I})$ and apply the EDM sampler Karras et al. (2022) with the same conditions over T steps.

We encode scene-level and text-level prompts separately and fuse them as

$$e_{\text{img}} = W_{\text{img}} \phi_{\text{img}}(\mathcal{S}_{\text{IMG}}), \quad e_{\text{txt}} = W_{\text{txt}} \phi_{\text{txt}} \left(\left\{ \left(\mathcal{S}_{\text{text}}^k, \mathcal{S}_{\text{pos}}^k \right) \right\}_{k=1}^K \right), \quad e^t = (1 - \lambda_t) e_{\text{txt}} + \lambda_t e_{\text{img}},$$
(4)

where $\phi_{\rm img}$ and $\phi_{\rm txt}$ are text encoders (kept frozen), and $W_{\rm img}, W_{\rm txt}$ are linear projections to a shared embedding space. The guided residual is computed via classifier-free guidance:

$$\widehat{\epsilon}_t = (1 + \omega) \mathcal{D}_{\theta}(z_t, t, e^t) - \omega \mathcal{D}_{\theta}(z_t, t, \varnothing), \tag{5}$$

with guidance scale ω . Consistently, the same λ_t also modulates residual injection (cf. Eq. 1) as a time-varying blend $\tilde{r}_l(t) = s_{\text{CTRL}} \left[(1 - \lambda_t) \, r_l^{\text{TXT}} + \lambda_t \, r_l^{\text{SR}} \right]$.

Binary Ping-Pong Policy. We found that a *binary* schedule that alternates between text-centric $(\lambda_t = 0)$ and image-centric $(\lambda_t = 1)$ guidance is effective:

$$\lambda_t = \begin{cases} 0, & \text{if } \left\lfloor \frac{t - t_0}{\tau} \right\rfloor \mod 2 = 0, \\ 1, & \text{otherwise,} \end{cases}$$
 (6)

where $\tau \in \mathbb{N}$ is the toggle period (default $\tau = 1$) and t_0 is an optional offset. Intuitively, the text-focused phases inject precise glyph cues, while the image-focused phases stabilize global structure and appearance. We also experimented with continuous ramps $\lambda_t = g(\sigma_t)$ (e.g., noise-level monotone schedules), but the square-wave "ping-pong" yielded the best OCR F_1 at similar perceptual quality (see 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 F₁ scores Chng et al. (2019) to quantify semantic correctness. Pixel-wise fidelity is measured by MANIQA Yang et al. (2022), CLIP-IQA Wang et al. (2023), and MUSIQ Ke et al. (2021) (see Sec. A.1). 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 LDM backbone and fine-tune it on our synthetic corpus generated with LLaVA-NeXT Liu et al. (2024), Nunchaku Cruanes et al. (2016), and SUPIR Yu et al. (2024b). Full data-generation pipelines and hyper-parameters and setup are detailed in Appendix B.

4.2 EVALUATION RESULTS

As shown in Table 1, many baseline methods improve Super-Resolution (SR) scores at the cost of Optical Character Recognition (OCR) performance. For instance, on SVT×4, **DiffBIR** achieves excellent SR metrics (47.82 MANIQA / 71.18 MUSIQ) but suffers from text hallucination, leading to a low OpenOCR F1 score of 38.73. Conversely, **StableSR** attains a high LLaVA-NeXT F1 (73.91) through conservative restoration, which results in a poor MUSIQ score of 24.44. This pattern repeats on SCUT-CTW1500×4. In contrast, our method consistently mitigates this trade-off. It achieves the best OpenOCR F1 score in 5/6 settings and the best GOT-OCR F1 in 4/6, all while maintaining top-tier SR performance. Notably, on SVT×8, it is the best across all six metrics, and on CUTE80×8, it leads all SR metrics while also securing the top OpenOCR F1 score (63.66).

Table 1: quantitative comparison of OCR F1-scores and SR quality metrics across datasets and models. red and blue indicate the best and second-best scores, respectively.

			OCR metric	F1	SR metric			
Dataset	Model	OpenOCR	GOT-OCR	LLaVA-NeXT	MANIQA	CLIP-IQA	MUSIQ	
	BSRGAN	53.96	58.66	68.50	38.16	39.63	66.25	
SVT (×4)	DiffBIR	38.73	42.33	45.19	47.82	58.66	71.18	
	DiffTSR	19.35	22.51	29.23	21.34	27.69	46.24	
	InvSR	57.79	60.96	65.00	46.78	57.30	70.81	
	PiSA-SR	63.30	65.23	67.75	37.41	44.30	61.87	
` ′	Real-ESRGAN	59.15	67.32	72.53	31.16	28.58	51.14	
	StableSR	59.88	63.76	73.91	24.75	32.18	24.44	
	SUPIR	58.41	61.90	62.14	42.36	48.42	67.55	
	GLYPH-SR (ours)	67.54	71.72	73.22	47.75	59.40	70.99	
	BSRGAN	24.67	21.86	35.10	51.41	47.44	67.52	
	DiffBIR	24.71	23.82	30.71	62.37	61.90	71.19	
	DiffTSR	19.77	15.98	23.69	35.39	30.59	55.83	
	InvSR	29.57	26.41	34.50	57.75	55.94	69.25	
SCUT-CTW1500 (×4)	PiSA-SR	37.46	34.14	44.11	56.31	53.05	68.19	
` ´	Real-ESRGAN	31.31	26.94	43.25	40.81	43.43	52.66	
	StableSR	25.55	19.95	45.86	31.04	43.61	24.92	
	SUPIR	18.26	17.61	24.37	57.35	51.68	66.96	
	GLYPH-SR (ours)	38.26	36.96	42.90	70.33	57.88	70.31	
	BSRGAN	73.09	56.02	83.97	44.22	55.73	69.13	
CUTE80 (×4)	DiffBIR	68.88	48.82	81.84	51.04	72.64	69.06	
	DiffTSR	61.08	47.48	73.71	33.94	38.47	58.74	
	InvSR	72.46	55.62	84.75	50.30	67.78	70.66	
	PiSA-SR	72.77	54.80	82.65	45.82	61.81	66.18	
	Real-ESRGAN	73.71	58.79	84.23	38.20	48.71	60.65	
	StableSR	72.14	57.22	82.92	36.26	49.74	60.09	
	SUPIR	70.85	51.87	82.11	47.50	62.62	68.26	
	GLYPH-SR (ours)	73.09	55.62	85.01	49.77	65.93	69.96	
	BSRGAN	14.61	13.12	25.56	37.14	37.58	62.83	
	DiffBIR	16.70	18.55	22.32	45.54	53.20	64.11	
	DiffTSR	10.28	10.72	15.87	21.39	26.39	43.96	
	InvSR	17.12	21.15	21.54	32.51	50.83	51.69	
$SVT(\times 8)$	PiSA-SR	17.53	24.05	37.76	34.02	18.39	30.24	
	Real-ESRGAN	17.73	23.29	30.83	28.38	17.86	43.01	
	StableSR	20.95	24.43	43.24	23.16	23.38	16.22	
	SUPIR	33.61	35.96	36.78	40.17	45.06	65.20	
	GLYPH-SR (ours)	48.79	56.16	58.54	47.40	56.78	69.93	
	BSRGAN	3.37	3.54	3.88	46.21	37.83	66.05	
	DiffBIR	4.76	5.10	4.64	54.75	49.89	63.16	
	DiffTSR	2.95	2.86	2.90	35.49	31.88	50.43	
	InvSR	2.09	2.17	2.43	29.65	29.62	40.29	
SCUT-CTW1500 (×8)	PiSA-SR	7.61	6.92	9.43	41.77	36.75	58.95	
` ´	Real-ESRGAN	5.02	5.64	7.74	28.37	20.95	39.99	
	StableSR	3.33	4.43	7.49	20.93	20.92	16.62	
	SUPIR	5.43	6.26	7.00	55.46	47.02	65.55	
	GLYPH-SR (ours)	11.09	14.71	14.67	61.94	48.21	63.43	
	BSRGAN	55.21	46.57	71.18	42.07	54.31	67.33	
	DiffBIR	59.56	44.71	70.53	47.53	62.09	64.62	
	DiffTSR	54.39	42.33	63.30	33.55	42.95	57.47	
	InvSR	56.42	45.18	72.46	37.66	62.43	57.69	
CUTE80 (×8)	PiSA-SR	52.72	42.33	75.24	30.71	30.80	45.16	
(-=/	Real-ESRGAN	59.18	49.27	74.33	35.17	36.46	56.55	
	StableSR	57.81	45.18	73.87	26.00	40.42	34.48	
	SUPIR	58.01	42.81	70.20	46.38	61.67	67.04	

Fig. 4 concretizes the two failure modes introduced earlier (Fig. 1). The examples on the left illustrate *hallucination*—sharp strokes that alter glyphs, raising IQA scores but breaking legibility. In contrast, those on the right exhibit *conservative restoration*. This issue stems from insufficient SR, a cautious approach to prevent hallucination. While this allows an OCR module to recognize the low-quality text, it results in blurry, low-contrast images with minimal SR gains. By preserving glyph topology while restoring realistic textures, **GLYPH-SR** avoids both pitfalls, yielding images that are both high-quality and OCR-readable. This outcome underscores why evaluations must report SR and OCR metrics jointly for a comprehensive assessment.



Figure 4: Qualitative examples illustrating the trade-off between SR metrics (e.g., MANIQA, CLIP-IQA, MUSIQ) and OCR metrics (F1, Accuracy) in scene-text images. While some methods improve perceptual SR scores, they may degrade OCR performance, and vice versa.

Superior OCR Fidelity. GLYPH-SR consistently achieves top-two F_1 scores across all datasets and OCR engines. On the most challenging benchmarks, it surpasses competitors by a large margin (e.g., +12.0 pp on CUTE80, \times 8), confirming the efficacy of our proposed token-wise guidance.

Competitive Perceptual Quality. While prioritizing text, GLYPH-SR maintains excellent global fidelity, ranking first or second in 26 out of 30 test cases across MANIQA, CLIP-IQA, and MUSIQ. It frequently outperforms other diffusion models like DiffBIR and SUPIR in these metrics.

Robustness Under Severe Degradation. The performance gap widens at $\times 8$ scale, where our model avoids the textual hallucination of GANs and the over-smoothing of generic diffusion methods. GLYPH-SR maintains high OCR scores without sacrificing perceptual quality, demonstrating its robustness to extreme degradation.

Taken together, the results confirm that our method yields a balanced architecture that advances the SOTA by resolving the conflict between text recognition and perceptual SR.



Figure 5: Comparison of SR results against different methods (DiffBIR, Real-ESRGAN, BSRGAN, and GLYPH-SR) on various degraded LR images.

Fig. 5 visually demonstrates how our model uniquely preserves text structure and legibility across severe degradations (×4 to ×8). Competing methods exhibit clear failure modes. Diffusion models like DiffBIR, despite high perceptual scores, frequently hallucinate incorrect characters (e.g., 'EANK OF ENUNAL'). Conversely, GAN-based methods like BSRGAN's high contrast produces jagged, geometrically distorted glyphs that harm human readability.

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 consistently reconciles both objectives, delivering coherent and legible results even at the extreme $\times 8$ scale where other models collapse.

4.2.1 ABLATION STUDIES



Figure 6: Four prompt settings using combinations of texts (S_{text}) and its spatial positions (S_{pos}).

Fig. 6 shows the effect of selectively removing the two of guidance used by GLYPH-SR: (i) the OCR string S_{text} and (ii) its spatial positions S_{pos} . We evaluate four combinations—both, text-only, position-only and none.

- 1) Full guidance ($S_{\text{text}} + S_{\text{pos}}$): The top-left quadrants reconstruct the text pattern without distortions, retaining stroke width, inter-letter spacing, and global geometry.
- 2) Text-only guidance ($\mathcal{S}_{text}/\mathcal{S}_{pos}$): When positional guidance is removed, the model hallucinates irregular kerning and warped baselines (e.g. "STASHOES COFFEE"), indicating that semantics alone cannot anchor glyph layout.
- 3) Position-only guidance (\mathcal{S}_{text} / \mathcal{S}_{pos}): Conversely, supplying bounding boxes but no textual content yields partial or incorrect spellings ("STABHOUES SOFFCE"), showing that location cues without semantics lead to character-level ambiguity.
- 4) No guidance ($\mathcal{S}_{text} + \mathcal{S}_{pos}$): Removing both priors produces the worst outcomes—severe hallucinations and geometric distortions reminiscent of generic diffusion SR.

5 CONCLUSIONS

Super-resolution research has traditionally prioritized perceptual quality, often neglecting a critical aspect of text-rich scenes: legibility. This creates a persistent gap where models produce sharp-looking images that still cannot be read correctly, as text is underweighted by standard SR objectives. To resolve this, GLYPH-SR reframes the task as a bi-objective problem that optimizes both visual realism and text legibility. We introduce a practical recipe featuring a VLM-guided diffusion model with a dual-branch TS-ControlNet, which fuses spatial OCR cues and a global caption. To properly evaluate this balance, we provide a factorized synthetic corpus and a dual-axis protocol pairing OCR F_1 with perceptual IQA metrics. On challenging benchmarks (SVT, SCUT-CTW1500, CUTE80 at $\times 4/\times 8$), GLYPH-SR improves OCR F_1 by up to +15.18 pp over strong baselines while maintaining top-tier perceptual quality. Future work will explore multilingual scripts, stronger geometric priors, and tighter integration with end-to-end recognition systems.

REPRODUCIBILITY STATEMENT

We are committed to the reproducibility of our work. The complete source code, pretrained models, synthetic data generation scripts, and evaluation suite for our **GLYPH-SR** framework are provided as supplementary material, with direct links available in Appendix B. Details of the model architectures, pre-trained backbones, and key hyper-parameters are described in Section 4.1 and extensively in Appendix B, which also specifies the hardware and software environment used for all experiments. Our dual-axis evaluation protocol, including all benchmark datasets (SVT, SCUT-CTW1500, CUTE80) and the specific OCR and perceptual metrics, is documented in Section 4.1 and Appendix A.1. The core components of our method, including the Text-SR Fusion ControlNet, condition decomposition, and the ping-pong scheduler, are detailed in Section 3. The data generation and fine-tuning workflow, serving as a practical guide, is outlined in Appendix B. Extended experimental results, comprehensive ablation studies, sensitivity analyses, and a discussion of the computational footprint are reported in Appendix C to ensure full transparency of our findings. These resources provide a comprehensive toolkit for the faithful reproduction and independent verification of our results.

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.
- 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.
- 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.
- 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 superresolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.
 - 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.
 - 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.

Supplementary Material

TABLE OF CONTENTS

A. Metric-based evaluation A.1. Evaluation metrics A.2. Misalignment between metrics and human perception B. Experiment detail B.1. Reproducibility Statement B.2. Setup B.2.1. Evaluation datasets B.2.2. Pre-trained models B.2.3. Key hyper-parameters B.2.4. Synthetic training data C. Experiment results C.1. Compare character generation to other models C.2. Sensitivity to upstream VLM/OCR errors C.3. Ablation study on ping-pong schedule C.4. Computational footprint and practical efficiency

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 performance, we report:

- F_I score, Precision, Recall and Accuracy (↑), : character-level measures of OCR correctness; higher is better.
- Normalized Edit Distance (1-NED) (†): inverse of edit distance, scaled to [0, 100]; higher values indicate closer agreement with the ground truth.

Image-Quality Metrics. For perceptual fidelity, we adopt:

- 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)) (\(\gamma\)): no-reference quality score based on attention-driven features, scaled to 0–100.
- CLIP-based Image Quality Assessment (CLIP-IQA Wang et al. (2023)) (†): semantic fidelity metric leveraging CLIP embeddings, 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 often drift from what people actually perceive—especially when the low-resolution input is heavily degraded. Fig. 7 offers four counter-examples that highlight three recurring failure modes.

Perceptual vs. Semantic Fidelity. In Figure 7, 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 7). 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, side-by-side inspection or user studies remain indispensable.



Figure 7: 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). The gap exposes a growing consensus: traditional metrics alone do not capture human perception of text-laden imagery.

810 EXPERIMENT DETAILS 811 812 **B.1** Reproducibility Statement 813 814 **Synth Dataset:** https://drive.google.com/drive/folders/1eYMvZQq-93okI2v1YldXLPHDycBkuvdu? 815 usp=drive link 816 817 **Pretrained Model:** 818 https://drive.google.com/drive/folders/1hrZ5jRbVLcRSFpbL-uPxe9iLddylAFgk? 819 usp=drive_link 820 821 https://drive.google.com/drive/folders/1A75nh0QEG1hcEhzUJxO75X8LfTO71R3K? 822 usp=drive_link 823 **Results:** 824 https://drive.google.com/drive/folders/1CArNuMOAI50z3TGsR66u218RLV5UdHYa? 825 usp=drive_link 826 827 Data Generation & Fine-Tuning Workflow 828 829 1. Stage 1 – Scene Description Extraction 830 dataset_generater/make_dataset_get_desc.py 831 ./datasets/descriptions/containing: {id, image_path, ocr_text, caption}. 832 2. Stage 2 – Augmented Prompt Synthesis 833 third_party/make_dataset_with_nunchaku/ 834 make_dataset_with_augmentation.py 835 Invokes the *Nunchaku* augmentation engine to expand each record with synthetic corruptions 836 (blur, noise, JPEG artifacts) and with diversity-enhanced textual prompts. The output is a 837 paired folder structure: ./datasets/aug/{hq, lq}. 838 3. Stage 3 – Negative/HQ Pairing 839 dataset_generater/make_dataset_Neg_HQ.py 840 Generates explicit (LQ, HQ) pairs and the associated prompt metadata required by GLYPH-841 SR. Final training files are placed under ./datasets/final/. 4. Stage 4 – Fine-Tuning 843 train_GLYPH_SR.py 844 845 python3 train_GLYPH_SR.py \ 846 --data_root ./datasets/ \ 847 GLYPH-SR/model_configs/model_config.yaml 848 849 Inference Workflow 850 851 1. Create the checkpoint directory. 852 Download every model file from the Pre-trained Checkpoints link and place them in a newly 853 created folder named ${\tt CKPT_PTH}$ at the project root. 854 2. Patch all path references. 855 Edit the three files listed below so that each points to the new directory, e.g. 856 CKPT_PTH/<checkpoint_name>.pth: 857 • GLYPH-SR/model_configs/model_config.yaml 858 • GLYPH-SR/run GLYPH_SR.py860 GLYPH-SR/CKPT_PTH.py 3. Run command. 862 Verify correct loading by launching a single-image run:

python3 run_GLYPH_SR.py --img_path ./image.jpg

Successful execution confirms that all checkpoints are discovered and that GLYPH-SR is ready for inference.

B.2 SETUP

All experiments were conducted on a workstation equipped with three NVIDIA RTX 6000 Ada GPUs (48 GB each), 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, low resolution, and noise. This diversity makes it well-suited for fine-tuning vision-language models to enhance OCR robustness. In our pipeline, ICDAR2017 is used to fine-tune LLaVA-NeXT, enabling it to better handle degraded scene-text images and produce more accurate token-level guidance.

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 nonlinear 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 low-resolution 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 1024×1024 resolution.

B.2.3 KEY HYPER-PARAMETERS

- Vision-Language Encoder. A frozen LLaVA-NeXT produces 2 048-dimensional multimodal embeddings that act as cross-attention keys; because the encoder is not fine-tuned, it adds zero trainable parameters.
- First Stage (VAE). A 256×256 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.



"<mark>id</mark>": "/SVT_image_x4/00_18.jpg"

"<mark>OCR</mark>": "Days Inn & Suites"

"prompt": "The image depicts a street scene with a focus on a sign for a hotel named **Days Inn & Suites.** The image has a casual, everyday quality to it, likely intended to show the location of the hotel for travelers or passersby."

.....



"<mark>id</mark>": "/SVT_image_x4/00_19.jpg"

"<mark>OCR</mark>": "Comfort Inn"

"prompt": "The image shows a sign for a hotel or motel named **Comfort Inn.** The sign is rectangular with rounded corners and is mounted on a vertical pole. The focus is on the sign, and the image is taken from a slightly lower angle, which makes the sign stand out against the sky."

Figure 8: Step 1: JSONL metadata—id, OCR text, and scene prompt—generated by LLAVA-NEXT.

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 8 illustrates the resulting metadata, produced by make_dataset_get_desc.py.

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. 9).



Figure 9: Step 2: Prompt augmentation with stylistic keywords to boost visual diversity.

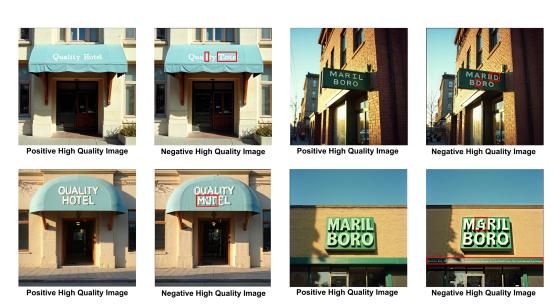


Figure 10: Step 3: Positive vs. intentionally corrupted (negative) high-quality pairs.

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. Corruptions are verified with the SUPIR pipeline Yu et al. (2024b)(Fig. 10).

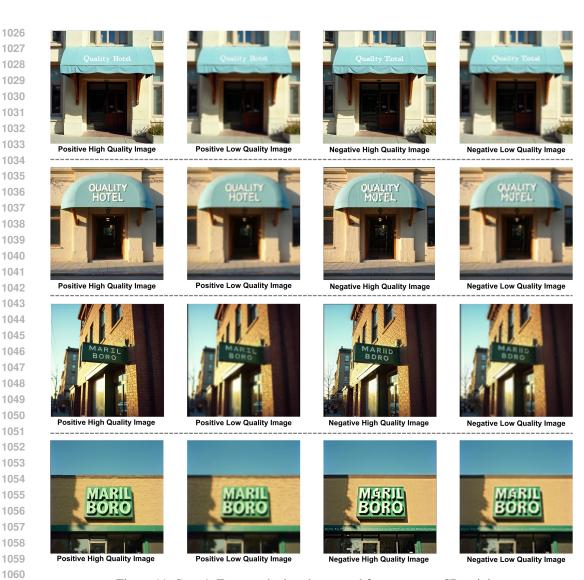


Figure 11: Step 4: Four synthetic subsets used for text-aware SR training.

Step 4: Final Dataset Assembly. All positive and negative images are merged into the four target splits shown in Fig. 11.

C EXPERIMENT RESULTS

C.1 COMPARE CHARACTER GENERATION TO OTHER MODELS.

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.

			OpenOCF	ł				GOT-OCE	ł			I	LaVA-Ne	KΤ	
	Precision	Recall	F_I score	1-NED	Accuracy	Precision	Recall	F_I score	1-NED	Accuracy	Precision	Recall	F_I score	1-NED	Accurac
SVT(x4)	1														
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
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
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
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)															
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
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
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
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)															
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
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
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
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)															
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
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
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
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)															
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
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
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
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)															
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
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
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
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

We compare our model's character generation ability against standard OCR models across difficulty levels. The results in Table 2 show significant improvements, especially under hard conditions. For evaluation, LLaVA-NeXT and our model were prompted using the following instructions: "Please perform OCR on this image." Additionally, both predicted and ground truth texts were normalized by removing non-alphabetic characters and ignoring case sensitivity before computing the metrics.

Table 3: Quantitative comparison of SR models on three scene-text datasets (SVT, SCUT-CTW1500, CUTE80) at $\times 4$ and $\times 8$ upscaling factors. Metrics include distortion-based (PSNR, SSIM, LPIPS \downarrow) 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
	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
SVT(x4)	InvSR	24.78	76.58	38.61	46.78	57.30	70.81
5 v 1 (x+)	PiSA-SR Real-ESRGAN	26.58 29.67	82.04 88.58	34.13 30.68	37.41 31.16	44.30 28.58	61.87 51.14
	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
	GLYPH-SR (ours)	22.89	67.19	42.20	47.75	59.40	70.99
	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
SCUT CTW/1500(4)	InvSR	18.32	60.71	32.99	57.75	55.94	69.25
SCUT-CTW1500(x4)	PiSA-SR	20.07	63.99	31.18	56.31	53.05	68.19
	Real-ESRGAN	20.85	67.46	36.81	40.81	43.43	52.66
	StableSR SUPIR	19.24 13.61	55.45 32.98	49.03 52.15	31.04 57.35	43.61 51.68	24.92 66.96
	GLYPH-SR (ours)	18.19	54.67	37.15	70.33	57.88	70.31
	BSRGAN	27.35	79.76	31.83	44.22	55.73	69.13
CUTE80(x4)	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
	Real-ESRGAN	28.14	82.30	32.01	38.20	48.71	60.65
	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
	GLYPH-SR (ours)	23.03	69.54	37.03	49.77	65.93	69.96
	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
SVT(x8)	InvSR	22.82	71.34	41.84	32.51	50.83	51.69
5 V 1 (X6)	PiSA-SR Real-ESRGAN	26.12 25.69	77.64 80.28	50.83 41.92	34.02 28.38	18.39 17.86	30.24 43.01
	StableSR	25.09 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
	GLYPH-SR (ours)	21.77	61.36	47.85	47.40	56.78	69.93
	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
COLUMN CORNEL SOOK ON	InvSR	11.81	30.68	65.88	29.65	29.62	40.29
SCUT-CTW1500(x8)	PiSA-SR	17.22	47.63	48.90	41.77	36.75	58.95
	Real-ESRGAN	17.65	52.34	52.14	28.37	20.95	39.99
	StableSR	17.00	43.50	66.02	20.93	20.92	16.62
	SUPIR	12.63	26.51	58.63 52.58	55.46	47.02 48.21	65.55
	GLYPH-SR (ours)	16.27	41.31		61.94		63.43
	BSRGAN DiffBIR	23.84 22.77	72.55 65.36	39.14 41.79	42.07 47.53	54.31 62.09	67.33 64.62
	DiffTSR	22.77	70.41	41.79	33.55	42.95	57.47
	InvSR	21.83	70.41	38.05	33.33 37.66	62.43	57.69
CUTE80(x8)	PiSA-SR	23.36	70.70	47.71	30.71	30.80	45.16
	Real-ESRGAN	24.01	75.58	39.60	35.17	36.46	56.55
	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
	GLYPH-SR (ours)	21.19	65.15	42.31	47.75	65.85	68.85

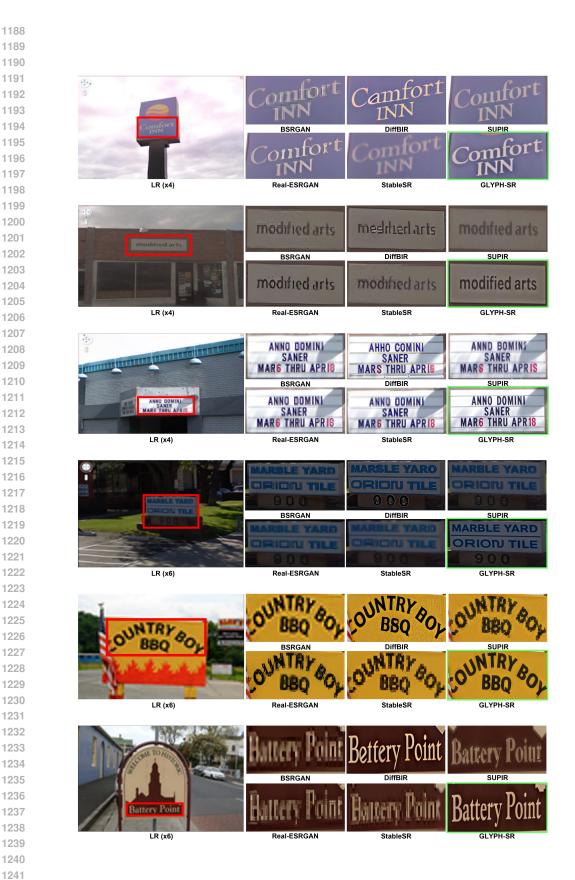




Figure 12: Qualitative comparison of scene-text SR under various degradation scales ($\times 4$, $\times 6$, $\times 8$). While prior methods often blur or hallucinate characters, **GLYPH-SR** accurately restores readable, coherent text. Zoom in for detail.

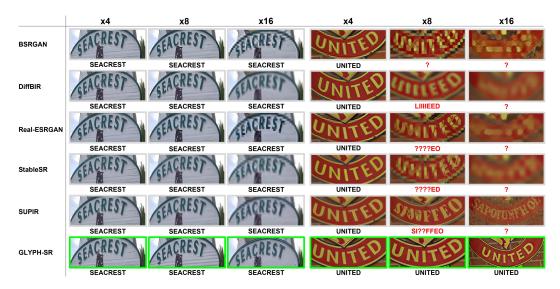


Figure 13: Qualitative comparison of text-centric SR results at $\times 4$, $\times 8$ and $\times 16$ scales.

Fig. 13 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 $\times 8$. GLYPH-SR achieves coherent integration of text and imagery even under $\times 16$ SR.

C.2 ABLATION STUDY ON PING-PONG SCHEDULE

Table 4 presents results for the CUTE80 benchmark at x4 and x8 scales under two evaluation

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

(a) CUTE80 (LR \times 4)

(b) CUTE80 (LR \times 8)

Scheduler Policy	MANIQA	CLIP-IQA	MUSIQ	OCR F_1	Scheduler Policy	MANIQA	CLIP-IQA	MUSIQ	OCR F_1
Binary ping-pong	49.77	65.93	69.96	85.01	Binary ping-pong	47.75	65.85	68.85	73.71
Mixing ($\lambda_t = 0.1$)	<u>49.95</u>	<u>70.64</u>	<u>70.67</u>	81.57	Mixing ($\lambda_t = 0.1$)	<u>48.89</u>	67.65	<u>69.56</u>	66.49
Mixing ($\lambda_t = 0.3$)	49.04	69.56	69.75	83.18	Mixing ($\lambda_t = 0.3$)	47.44	<u>68.31</u>	68.86	69.87
Mixing ($\lambda_t = 0.5$)	47.57	65.47	68.95	84.23	Mixing ($\lambda_t = 0.5$)	46.57	64.07	67.35	73.40
Mixing ($\lambda_t = 0.7$)	47.86	68.91	68.83	81.84	Mixing ($\lambda_t = 0.7$)	45.80	67.98	67.19	66.84
Mixing ($\lambda_t = 0.9$)	48.85	69.11	69.13	82.65	Mixing ($\lambda_t = 0.9$)	45.58	67.66	67.18	68.88

protocols. The binary strategy yields higher CLIP-IQA and MUSIQ scores—reflecting superior perceptual quality—while simultaneously boosting the OCR F_I score (LLaVA-NeXT), supporting its effectiveness at balancing text readability and image fidelity.

C.3 SENSITIVITY TO UPSTREAM VLM/OCR ERRORS

We assess how errors in upstream text guidance (OCR/VLM) propagate to GLYPH–SR. Because our method deliberately conditions on tokenlevel strings and locations, corrupted guidance could degrade both readability and overall perceptual quality. We simulate three error modes and measure their impact on OCR and IQA metrics.

- 1. Random Character Corruption: Replace $n\% \in \{30, 50, 90\}$ of characters in the OCR string with uniformly sampled alternatives (random noise).
- 2. **Plausible Character Swaps ("Swap"):** Systematically replace characters with visually confusable counterparts from a curated set (e.g., $0 \leftrightarrow 0$, $1 \leftrightarrow 1$, $1 \leftrightarrow 7$).
- 3. **Missed Detections** ("**Drop**"): Remove a portion of OCRrecognized characters to emulate detection/recognition failures.

Table 5 reports OpenOCR/GOTOCR F_1 and MANIQA/CLIPIQA. Parentheses show absolute changes w.r.t. the uncorrupted baseline.

Table 5: **Sensitivity to OCR/VLM guidance errors.** Values in parentheses are absolute deltas from the baseline (lower is worse).

Error rate / Type	OpenOCR F_1	GOT-OCR F_1	MANIQA	CLIP-IQA
Baseline	48.82	38.36	62.01	79.69
30%	38.36 (-10.46)	28.67 (-9.69)	45.87 (-16.14)	63.65 (-16.04)
50%	32.03 (-16.79)	26.35 (-12.01)	45.39 (-16.62)	64.88 (-14.81)
90%	27.52 (-21.30)	26.35 (-12.01)	45.61 (-16.40)	66.00 (-13.59)
Swap	39.88 (-8.94)	33.12 (-5.24)	45.81 (-16.20)	66.00 (-13.69)
Drop	41.85 (-6.97)	32.03 (-6.33)	44.82 (-17.19)	65.30 (-14.39)

All error modes substantially hurt both axes: readability ($OpenOCR/GOTOCR\ F_1$) and perceived image quality (MANIQA/CLIPIQA). Even moderate noise (50%) reduces OpenOCR F_1 by 16.79 pp and MANIQA by 16.62 points. Plausible swaps and missed detections also trigger large drops, indicating that *quantity* (how many tokens are wrong), *nature* (plausible vs. random), and *absence* (drops) all impair glyph integrity and global appearance. This validates our design choice to use a strong, LRaware OCR/VLM and to treat guidance quality as a firstorder factor in textaware SR.



Figure 14: Failure cases where GLYPH-SR produces visually plausible SR outputs but incorrectly generates text in non-textual regions.

As illustrated in Fig. 14, GLYPH-SR can deliver visually plausible SR results yet still *hallucinates* glyphs in regions that were originally non-textual. This deficiency in text-region localization means the reconstructed text may be ambiguous, incomplete, or entirely spurious. Furthermore, when multiple words are present, the model tends to enhance only the most visually salient word and overlook the rest. These failure cases underline the necessity for finer-grained attention mechanisms and explicit supervision of glyph positions in future work.

C.4 COMPUTATIONAL FOOTPRINT AND PRACTICAL EFFICIENCY

Setup. We benchmark inference on a $4 \times SR$ task with 512×512 inputs. Times are mean \pm 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.

Table 6: **Compute comparison.** For VLMguided methods, #Params lists (*restoration*, *VLM*) in millions.

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

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 lowresolution 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 generalpurpose VLM with a compact, LRtextspecialized 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 crossattention KV maps), skipping redundant compute and lowering endtoend runtime.

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

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** (\approx 54.8M parameters) that handles text-guidance fusion.
- 2. VLM LoRA adapter (\approx 5.9M parameters) with low rank (r=8), lora_alpha of 32, and dropout of 0.05.

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

Table 7: **Trainable parameter counts** (millions). Despite using a VLM, GLYPH–SR keeps *trainable* parameters modest via freezing and LoRA.

Metric	GLYPH-SR	PiSA-SR	SeeSR	StableSR	DiffBIR	SUPIR	DiffTSR
Trainable (M)	60.7	0.38	489.04	152.67	378.95	3865.64	55.31

Inference latency. The OCR/VLM guider is the main overhead driver. Out of a total per-image latency of 38.25 ± 1.28 seconds (Sec. C.4), the VLM component accounts for ≈ 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 6): 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).

REFERENCES FOR APPENDIX

- [12] 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), pages 2286–2295, 2022.
- [13] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In AAAI, 2023.
- [14] 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), pages 8827–8837, 2021.
- [18] 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.
- [20] 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, pages 25669–25680, 2024.
- [21] 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 2018.
- [22] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
- [23] 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, pages 633–651. Springer, 2020.
- [24] Jinjin Gu, Haoming Cai, Chao Dong, Jimmy S. Ren, Radu Timofte, Yuan Gong, Shanshan Lao, Shuwei Shi, Jiahao Wang, Sidi Yang, Tianhe Wu, Weihao Xia, Yujiu Yang, Mingdeng Cao, Cong Heng, Lingzhi Fu, Rongyu Zhang, Yusheng Zhang, Hao Wang, Hongjian Song, Jing Wang, Haotian Fan, Xiaoxia Hou, Ming Sun, Mading Li, Kai Zhao, Kun Yuan, Zishang Kong, Mingda Wu, Chuanchuan Zheng, Marcos V. Conde, Maxime Burchi, Longtao Feng, Tao Zhang, Yang Li, Jingwen Xu, Haiqiang Wang, Yiting Liao, Junlin Li, Kele Xu, Tao Sun, Yunsheng Xiong, Abhisek Keshari, Komal, Sadbhawana Thakur, Vinit Jakhetiya, Badri N Subudhi, Hao-Hsiang Yang, Hua-En Chang, Zhi-Kai Huang, Wei-Ting Chen, Sy-Yen Kuo, Saikat Dutta, Sourya Dipta Das, Nisarg A. Shah, and Anil Kumar Tiwari. Ntire 2022 challenge on perceptual image quality assessment. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 951–967, June 2022.
- [25] 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), pages 25669–25680, June 2024.