

TOLA: Text-aware One-Step Latent Adaptation for Diffusion-based Text Image Super-Resolution

Yike Xu^{1,2*} Yue Shi^{1,3} Yong Guo¹ Jiezhong Cao^{1†}

¹Shanghai Jiao Tong University ²University of Chinese Academy of Sciences ³Shanghai AI Laboratory

Abstract

Text image super-resolution (TSR) aims to recover visually faithful and readable text under unknown degradations. Existing diffusion-based methods typically rely on multi-step prediction of either the high-resolution image or its text prior, resulting in prohibitive computational cost and inference latency. More critically, an erroneous text prior may be repeatedly injected into the denoising process, causing image and text predictions to reinforce each other and progressively amplify an early recognition error into a sharp yet semantically incorrect character. To address these limitations, we propose TOLA, a Text-aware One-step Latent Adaptation framework without iterative image-text diffusion. TOLA consists of two key modules. First, a confidence-weighted text conditioning module constructs the semantic condition only once and suppresses unreliable OCR predictions before they contaminate image reconstruction. Second, a lightweight latent residual correction module explicitly estimates and corrects the structured residual errors to recover missing or distorted stroke details. Extensive experiments demonstrate our state-of-the-art performance across all evaluation metrics on both CTR-TSR-Test ($\times 4$) and RealCE-200 benchmarks. It is worth noting that our TOLA consistently surpasses existing diffusion-based TSR methods by at least 2.72 dB in PSNR on CTR-TSR-Test.

Introduction

Text image super-resolution (TSR) aims to recover high-resolution text images from degraded low-resolution observations while preserving both visual details and character-level semantics (Zhao et al., 2021). This challenge becomes more pronounced in the blind setting, where the degradation process is unknown. Unlike natural-image super-resolution, TSR must preserve not only perceptual quality but also character-level semantic accuracy, as a single missing, merged, or hallucinated stroke may lead to an incorrect transcription even when the reconstructed image appears visually sharp. There-

fore, TSR models need to recover fine-grained visual structures while maintaining semantic fidelity.

Early TSR methods enhance reconstruction by incorporating sequence modeling, text-aware attention, recognition guidance, and character-structure priors (Wang et al., 2020; Ma et al., 2022, 2023a; Li et al., 2023). These methods perform well when the low-resolution input still contains reliable structural or semantic cues. Under severe degradations, however, missing and merged strokes make both visual reconstruction and text recognition highly ambiguous. Diffusion models have recently emerged as a promising solution because their generative priors can synthesize plausible high-frequency details that deterministic restoration networks often fail to recover (Ho et al., 2020; Rombach et al., 2022; Kavar et al., 2022; Saharia et al., 2023; Wang et al., 2024a). Existing diffusion-based TSR methods typically rely on multi-step prediction of either the high-resolution image or its text-aware prior (Zhang et al., 2024; Xu et al., 2026), as shown in figure 1. Although iterative refinement improves generative capacity, it also introduces two fundamental limitations. First, repeatedly evaluating image and text networks incurs computational cost and inference latency. Second, the text prior used to guide reconstruction is itself inferred from the degraded observation and may therefore be incorrect (Yang et al., 2024). When an erroneous prior is repeatedly injected into the denoising process, image and text predictions may reinforce each other, progressively amplifying an early recognition error into a sharp yet semantically incorrect character.

Directly reducing multi-step diffusion-based TSR to a single image-denoising step is non-trivial due to two challenges. First, reliable text guidance must be constructed in a single pass without repeatedly injecting uncertain OCR predictions into image reconstruction. Second, the clean latent estimated from one high-noise prediction inevitably contains structured errors, especially around thin strokes, character boundaries, and ambiguous local patterns, which can no longer be corrected by subsequent reverse updates. Introducing another iterative prior-recovery or refinement process would merely shift, rather than eliminate, the computational burden. The key

*Work done during an internship at Shanghai Jiao Tong University.

†Corresponding author.

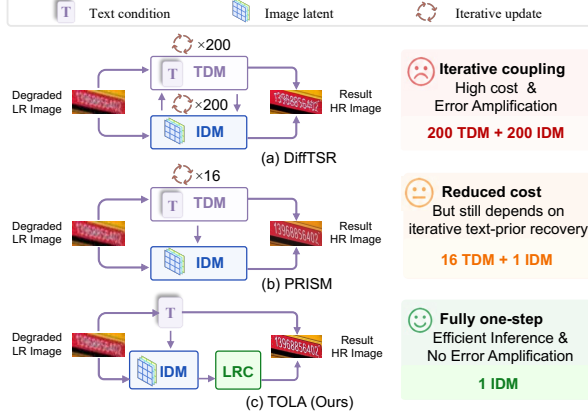


Figure 1: Comparison of three different diffusion-based TSR inference paradigms. (a) DiffTSR performs 200-step iterative text-prior and image restoration. (b) PRISM uses one-step image reconstruction with 16-step text-prior recovery. (c) TOLA constructs the text condition once, evaluates IDM once, and applies a Latent Residual Correction (LRC) for clean-latent correction. Prior iterative text-image refinement may repeatedly propagate inaccurate semantic cues through multiple denoising stages, leading to accumulated recognition errors. In contrast, TOLA avoids error amplification by designing a one-step text-conditioned IDM.

is therefore to combine one-time semantic conditioning with explicit, non-iterative latent residual correction.

Based on this observation, we propose TOLA, a Text-aware One-step Latent Adaptation framework. Our central insight is that semantic conditioning and latent residual correction need not be coupled along the same reverse trajectory. Instead, reliable semantic evidence can be selected once before image prediction, while structured reconstruction errors can be modeled explicitly afterward in the clean-latent space. As illustrated in figure 2, TOLA consists of two complementary modules. First, the confidence-weighted text conditioning module uses frozen TransOCR to predict text tokens and token-wise confidence scores, then constructs the semantic condition once via the frozen MoM module. By suppressing uncertain OCR predictions, this module reduces the risk that an unreliable text hypothesis will dominate image reconstruction. Second, the lightweight latent residual correction module explicitly estimates the structured residual remaining in the initial clean latent and restores missing or distorted stroke details without introducing another reverse trajectory.

Our main contributions are summarized as follows:

- We propose TOLA, which reformulates the original multi-step image-text diffusion pipeline into a non-iterative framework with one-shot text conditioning and one-step latent prediction.
- We introduce confidence-weighted text condition-

ing, which constructs the semantic condition once and suppresses unreliable character priors using OCR confidence scores. This prevents erroneous text information from being repeatedly injected and amplified during denoising, yielding reliability-aware semantic guidance.

- Extensive experiments demonstrate that TOLA achieves state-of-the-art performance across all reported metrics on CTR-TSR-Test ($\times 4$) and RealCE-200. TOLA outperforms the official 200-step DiffTSR in all nine metrics, improving PSNR by 4.2137 dB while reducing FID by 12.2271. Meanwhile, it shortens the average inference time from 9.9548s to 0.1319s, achieving a $75.5\times$ speedup.

Related Work

Text-Aware Scene Text Image Super-Resolution.

Generic SR optimizes visual fidelity without preserving character identity (Dong et al., 2016; Wang et al., 2018; Chen et al., 2022b; Liang et al., 2021). TSR models strokes, sequences, layout, deformation, and location (Wang et al., 2020; Chen et al., 2021; Ma et al., 2022; Chen et al., 2022a; Zhao et al., 2022; Zhu et al., 2023b; Guo et al., 2023; Wei et al., 2025; TomyEnrique et al., 2024). Recognition, style, and structure priors provide guidance (Ma et al., 2023a; Li et al., 2023; Zhu et al., 2023a; Park and Ko, 2025; Yuan et al., 2025), but cues extracted from degraded inputs remain unreliable (Kong et al., 2024).

Diffusion-Based Text Image Restoration. Diffusion TSR conditions denoising on masks, LR text, recognition, or character embeddings (Liu et al., 2025; Zhou et al., 2024; Noguchi et al., 2024; Singh et al., 2024). DiffTSR couples IDM and TDM through MoM; Boosting DiffTSR adds mixed training, ResShift, cross-attention, and confidence weighting (Zhang et al., 2024; Pan et al., 2025). TextSR, TEXTS-Diff, and TADiSR address multilingual or real-world restoration (Ye et al., 2025; He et al., 2026; Hu et al., 2025); DualTSR and TeReDiff combine text diffusion or recognition with restoration (Niu et al., 2026; Min et al., 2026). These methods retain iterative refinement.

One-Step Diffusion Restoration. DDIM, ResShift, SinSR, and AddSR shorten or distill diffusion trajectories (Song et al., 2021; Yue et al., 2023; Wang et al., 2024b; Tai et al., 2026; Dong et al., 2025); OSediff uses LoRA and score distillation (Hu et al., 2022; Wu et al., 2024), while FiDeSR combines frequency injection with latent residual refinement (Kim et al., 2026). These methods target generic SR rather than the coupled DiffTSR process. PRISM is text-specific but uses $K = 16$ Euler steps for prior recovery (Xu et al., 2026). TOLA instead adapts pretrained DiffTSR to one IDM evaluation by con-

structing confidence-weighted TransOCR–MoM conditioning once and correcting the latent after noise-to-clean conversion.

Method

Overview

As shown in figure 2, TOLA converts iterative DiffTSR into a single-pass restoration path. TransOCR predicts tokens and confidence scores, while the LoRA-adapted VAE encoder maps the LR image to $\mathbf{z}_{\text{LR}}$ and q-sampling obtains $\mathbf{z}_t$ at $t = 999$. MoM fuses confidence-weighted token embeddings with $[\mathbf{z}_{\text{LR}}, \mathbf{z}_t]$ to construct $\mathbf{C}_{\text{cond}}$ once. Conditioned on $\mathbf{C}_{\text{cond}}$, the LoRA-adapted IDM performs one noise prediction, which is converted into the initial clean latent $\hat{\mathbf{z}}_0$. LRC then corrects the residual $\mathbf{z}_{\text{LR}} - \hat{\mathbf{z}}_0$ before VAE decoding. Only LoRA and LRC are optimized, and inference requires neither TDM nor iterative reverse updates.

Text-Aware One-Step Adaptation

Fixed-Timestep q-Sampling. The LoRA-adapted VAE encoder first encodes the degraded input image $\mathbf{x}_{\text{LR}}$ into an unscaled latent representation $\mathbf{z}_{\text{LR}}$,

$$\mathbf{z}_{\text{LR}} = \mathcal{E}_{\text{VAE}}(\mathbf{x}_{\text{LR}}). \quad (1)$$

Following DiffTSR q-sampling, we sample a noisy latent at the fixed, zero-indexed timestep $t = 999$:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t} \mathbf{z}_{\text{LR}} + \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (2)$$

where $\bar{\alpha}_t = \prod_{\tau=0}^t (1 - \beta_\tau)$ follows the 1000-step linear schedule, with $\beta_0 = 0.0015$ and $\beta_{999} = 0.0205$. We draw one Gaussian sample per input and supply $\mathbf{z}_{\text{LR}}$ and $\mathbf{z}_t$ to IDM on the VAE latent scale. Unlike TOLA, OSEDIff feeds the unnoised LR latent to its one-step U-Net (Wu et al., 2024).

Confidence-Weighted Text Conditioning. The frozen TransOCR instance used for conditioning predicts a token sequence and the corresponding token-wise confidence scores,

$$\begin{aligned} (\tilde{\mathbf{c}}^{\text{nat}}, \boldsymbol{\alpha}) &= \text{TransOCR}(\mathbf{x}_{\text{LR}}), \\ \tilde{\mathbf{c}} &= \text{Map}(\text{Norm}(\tilde{\mathbf{c}}^{\text{nat}})). \end{aligned} \quad (3)$$

Norm normalizes characters to Simplified Chinese; Map maps them to the IDM vocabulary with $\boldsymbol{\alpha}$ alignment. Following Boosting DiffTSR (Pan et al., 2025), confidence-weighted embeddings enter MoM, which computes

$$[\mathbf{I}_{\text{cond}}, \mathbf{C}_{\text{cond}}] = \mathcal{M}(s[\mathbf{z}_{\text{LR}}, \mathbf{z}_t], \boldsymbol{\alpha} \odot \text{Emb}(\tilde{\mathbf{c}}), t), \quad (4)$$

Algorithm 1 TOLA one-step inference.

Require: LR image $\mathbf{x}_{\text{LR}}$; fixed $t = 999$

Output: Restored image $\mathbf{x}_{\text{SR}}$

- 1: $\mathbf{z}_{\text{LR}} \leftarrow \mathcal{E}_{\text{VAE}}(\mathbf{x}_{\text{LR}}); (\tilde{\mathbf{c}}^{\text{nat}}, \boldsymbol{\alpha}) \leftarrow \text{TransOCR}(\mathbf{x}_{\text{LR}})$
 - 2: $\tilde{\mathbf{c}} \leftarrow \text{Map}(\text{Norm}(\tilde{\mathbf{c}}^{\text{nat}}))$
 - 3: Draw $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$; construct $\mathbf{z}_t$ with equation 2
 - 4: $[\mathbf{I}_{\text{cond}}, \mathbf{C}_{\text{cond}}] \leftarrow \mathcal{M}(s[\mathbf{z}_{\text{LR}}, \mathbf{z}_t], \boldsymbol{\alpha} \odot \text{Emb}(\tilde{\mathbf{c}}), t)$
 - 5: $\hat{\boldsymbol{\epsilon}}_\theta \leftarrow f_\theta([\mathbf{z}_t, \mathbf{z}_{\text{LR}}], t, \mathbf{C}_{\text{cond}})$
 - 6: Reconstruct $\hat{\mathbf{z}}_0$ with equation 6
 - 7: $\mathbf{r} \leftarrow \mathbf{z}_{\text{LR}} - \hat{\mathbf{z}}_0; \Delta \mathbf{r} \leftarrow g_\phi([\mathbf{z}_{\text{LR}}, \mathbf{r}])$
 - 8: $\hat{\mathbf{z}}_0^{\text{corr}} \leftarrow \mathbf{z}_{\text{LR}} - (\mathbf{r} + \Delta \mathbf{r}); \mathbf{x}_{\text{SR}} \leftarrow \mathcal{D}_{\text{VAE}}(s^{-1} \hat{\mathbf{z}}_0^{\text{corr}})$
-

where $s = 0.18215$ scales image latents and $\odot$ broadcasts token confidences over embeddings. IDM uses only $\mathbf{C}_{\text{cond}}$ for cross-attention; $\mathbf{I}_{\text{cond}}$ is discarded. Confidence is neither calibrated nor thresholded, and MoM runs once.

One-Step Clean-Latent Reconstruction. The LoRA-adapted IDM receives channel-wise concatenated noisy and LR latents and predicts

$$\hat{\boldsymbol{\epsilon}}_\theta = f_\theta([\mathbf{z}_t, \mathbf{z}_{\text{LR}}], t, \mathbf{C}_{\text{cond}}), \quad (5)$$

where $[\cdot, \cdot]$ is channel-wise concatenation and θ combines frozen base and trainable LoRA parameters. Since DiffTSR predicts noise, the clean latent is

$$\hat{\mathbf{z}}_0 = \frac{\mathbf{z}_t - \sqrt{1 - \bar{\alpha}_t} \hat{\boldsymbol{\epsilon}}_\theta}{\sqrt{\bar{\alpha}_t}}. \quad (6)$$

No reverse update follows equation 6; $\hat{\mathbf{z}}_0$ is passed directly to LRC. Algorithm 1 summarizes inference.

TOLA evaluates the VAE encoder/decoder, TransOCR, MoM, IDM, and LRC once each while omitting TDM; DiffTSR evaluates IDM and TDM 200 times each.

Latent Residual Correction

A FiDeSR-inspired LRC corrects local errors after noise-to-clean conversion (Kim et al., 2026). FiDeSR refines a directly predicted residual, whereas TOLA corrects the clean-latent residual without changing semantic conditioning.

Given the initial clean-latent estimate $\hat{\mathbf{z}}_0$, we define

$$\mathbf{r} = \mathbf{z}_{\text{LR}} - \hat{\mathbf{z}}_0. \quad (7)$$

From the concatenated input $[\mathbf{z}_{\text{LR}}, \mathbf{r}]$, LRC predicts

$$\Delta \mathbf{r} = g_\phi([\mathbf{z}_{\text{LR}}, \mathbf{r}]), \quad (8)$$

The trainable g_ϕ uses a 1×1 projection from 6 to 32 channels, one RRDB of three five-layer RDBs with 3×3 convolutions, and a 1×1 projection to 3 channels, with growth width 16 and residual scaling 0.2.

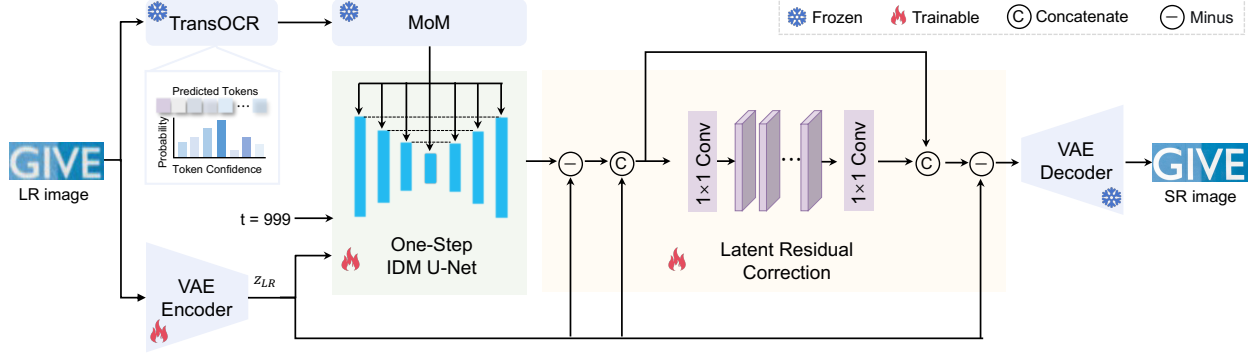


Figure 2: TOLA architecture. The LR image is encoded and q-sampled at fixed $t = 999$, while confidence-weighted TransOCR tokens and latent features form one MoM condition. A LoRA-adapted IDM predicts noise once, and LRC corrects the clean latent before frozen VAE decoding. Snowflakes and flames denote frozen and trainable components, respectively.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FID $\downarrow$	T-ACC $\uparrow$	T-NED $\uparrow$	P-ACC $\uparrow$	P-NED $\uparrow$
SRCNN	22.0288	0.5860	0.6021	0.3886	156.9406	0.4255	0.6107	0.3303	0.5279
ESRGAN	21.7741	0.5588	0.6455	0.3979	155.7055	0.4188	0.6061	0.3287	0.5273
NAFNet	21.5568	0.5678	0.5924	0.3776	144.9606	0.4187	0.5998	0.2998	0.4870
TSRN	20.4381	0.5840	0.6768	0.3926	140.2070	0.3307	0.5192	0.2238	0.4196
TBSRN	20.6883	0.5840	0.6510	0.3918	137.1033	0.3539	0.5425	0.2574	0.4563
TATT	20.9523	0.5846	0.6614	0.3896	125.6047	0.3558	0.5450	0.2491	0.4466
MARCONet	21.3460	0.6132	0.4386	0.3566	102.7384	0.3895	0.5735	0.2796	0.4600
PRISM	22.08--	--	0.2314	--	12.57--	0.4212	0.6644	--	--
DiffTSR	20.5958	0.6261	0.3103	0.2428	24.9120	0.4645	0.6519	0.3970	0.5873
TeReDiff	19.6333	0.5125	0.4714	0.3466	77.0387	0.2893	0.4491	0.1972	0.3592
Ours	24.8095	0.7235	0.2288	0.1953	12.6849	0.5401	0.7370	0.4626	0.6585

Table 1: Results on 8,089 CTR-TSR-Test $\times 4$ images. Non-diffusion methods are retrained on the CTR-TSR split, and all unified-protocol outputs share one evaluation implementation. T/P denote TransOCR/PaddleOCR; bold marks the best unified-protocol result. PRISM (Xu et al., 2026) is a BTL-test $\times 4$ /PP-OCRv5 cross-protocol reference; dashes preserve its reported precision. *Note:* “--” denotes unavailable values. *PRISM results are from its arXiv paper due to unavailable code.*

The corrected residual and latent are

$$\mathbf{r}^{\text{corr}} = \mathbf{r} + \Delta \mathbf{r}, \quad \hat{\mathbf{z}}_0^{\text{corr}} = \mathbf{z}_{\text{LR}} - \mathbf{r}^{\text{corr}}. \quad (9)$$

The frozen VAE decoder then reconstructs

$$\mathbf{x}_{\text{SR}} = \mathcal{D}_{\text{VAE}}(s^{-1} \hat{\mathbf{z}}_0^{\text{corr}}), \quad (10)$$

The inherited scale is $s = 0.18215$. Zero-initializing the LRC output projection initially preserves $\hat{\mathbf{z}}_0$ and learns only the residual correction.

Training Objective

TOLA uses pixel, perceptual, and recognition objectives:

$$\mathcal{L} = \mathcal{L}_1(\mathbf{x}_{\text{SR}}, \mathbf{x}_{\text{HR}}) + \mathcal{L}_{\text{LPIPS}}^{\text{VGG}}(\mathbf{x}_{\text{SR}}, \mathbf{x}_{\text{HR}}) + 0.02 \mathcal{L}_{\text{OCR}}^{\text{CE}}(\text{OCR}(\mathbf{x}_{\text{SR}}), \mathbf{y}_{\text{GT}}). \quad (11)$$

$\mathbf{x}_{\text{HR}}$ and $\mathbf{y}_{\text{GT}}$ are the ground-truth image and transcription; $\mathcal{L}_1$ is ℓ_1 , $\mathcal{L}_{\text{LPIPS}}^{\text{VGG}}$ is LPIPS-VGG, and $\mathcal{L}_{\text{OCR}}^{\text{CE}}$ is token-level cross-entropy. Two frozen TransOCR instances provide conditioning and OCR supervision; OCR gradients

update LoRA and LRC. Conditioning uses the IDM vocabulary, whereas $\mathcal{L}_{\text{OCR}}^{\text{CE}}$ retains the native TransOCR labels. No latent ℓ_1 , noise-prediction, distillation, or teacher loss is used.

Experiments

Experimental Setup

Datasets. We use 62,371 training and 1,273 validation images from the CTR-TSR preprocessing set (Zhang et al., 2024). LR inputs are generated online with Real-ESRGAN or BSRGAN degradations at scales $\{1, 2, 4\}$ (Wang et al., 2021; Zhang et al., 2021); all images use a 128×512 canvas. Evaluation uses all 8,089 pairs in CTR-TSR-Test $\times 4$.

For real-image generalization, we use two complementary test sets. RealCE-200 contains 200 filtered and deduplicated pairs from the RealCE validation split (Ma et al., 2023b), using matched official 13 mm and 52 mm crops as LR and reference images. We construct RT50

to extend text-category and acquisition-condition coverage beyond existing paired benchmarks. Its 50 images, collected through multiple routes, include 17 Chinese, 17 English, and 16 numeric samples spanning varied blur, illumination, viewpoints, resolutions, and backgrounds. A predefined output-independent protocol screens content validity, geometry, degradation difficulty, quality, and duplication. Neither set is used for training, validation, or checkpoint selection.

Evaluation Protocol and Metrics. To ensure a fair comparison, we adapt all non-diffusion methods to $\times 4$ restoration and retrain them on the same CTR-TSR training split. All unified-protocol methods are evaluated on the same 8,089-pair manifest at 128×512 using identical filename matching and metric implementations. We report luminance-channel PSNR and SSIM without border cropping (Wang et al., 2004), LPIPS-AlexNet (Zhang et al., 2018), DISTS (Ding et al., 2022), and FID (Heusel et al., 2017); training instead uses LPIPS-VGG. TransOCR predictions undergo full-to-half-width and simplified-Chinese conversion and whitespace removal while preserving case. ACC denotes exact-match accuracy, and NED denotes mean normalized edit-distance similarity. Independent PaddleOCR ACC and NED (Du et al., 2020) provide a recognizer cross-check.

Compared Methods. We compare TOLA with SRCNN, ESRGAN, NAFNet, TSRN, TBSRN, TATT, MARCONet, DiffTSR, and TeReDiff, and additionally include the published PRISM results as a cross-protocol reference.

Implementation Details. Starting from the official CTR-trained DiffTSR checkpoint, we optimize rank-4 LoRA and LRC for 100K steps on four NVIDIA RTX PRO 6000 GPUs with batch size 16 per GPU. FP32 AdamW uses learning rate 5×10^{-5} , $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.01, 500-step linear warm-up, and gradient clipping at 1.0. Training takes about 55 hours, and validation selects the checkpoint. Loss weights are 1, 1, and 0.02. Each image is restored once without sample selection or averaging.

Diagnostic Protocol. All trainable variants in Table 4 share CTR data, DiffTSR initialization, data order, optimization, and the 8,089-image evaluation. Conditioning controls alter only the fixed checkpoint input; GT prior is an oracle. Paired comparisons use identical per-image noise.

Main Results

Quantitative Comparisons. Table 1 shows that TOLA surpasses DiffTSR in all nine metrics, gaining 4.2137 dB

Method	PSNR $\uparrow$	SSIM $\uparrow$	T-ACC $\uparrow$	T-NED $\uparrow$	P-ACC $\uparrow$	P-NED $\uparrow$
SRCNN	19.1609	0.6024	0.3850	0.5837	0.4950	0.6670
ESRGAN	19.2062	0.5971	0.3850	0.5799	0.5000	0.6710
NAFNet	18.9038	0.5920	0.3650	0.5704	0.4750	0.6406
TSRN	18.6767	0.6009	0.2850	0.5103	0.3850	0.5649
TBSRN	18.9903	0.5947	0.3400	0.5364	0.3950	0.5980
TATT	18.9725	0.5954	0.3300	0.5291	0.4000	0.5970
MARCONet	18.3506	0.6198	0.3500	0.5408	0.3800	0.5972
DiffTSR	17.6287	0.5970	0.3500	0.5402	0.4300	0.6188
TeReDiff	17.8180	0.5272	0.2900	0.4826	0.3350	0.5508
Ours	19.2240	0.6199	0.3950	0.5841	0.5100	0.6720

Table 2: RealCE-200 cross-dataset results on the same 200 pairs. RealCE is excluded from training and checkpoint selection; T/P denote TransOCR/PaddleOCR.

PSNR and 0.0756 TransOCR ACC while reducing FID by 12.2271. PaddleOCR confirms this trend: ACC/NED rise from 0.3970/0.5873 to 0.4626/0.6585.

CTR-TSR Qualitative Results. Figure 3 shows more coherent strokes and closer character shapes and layouts across Chinese, English, and numeric examples.

Cross-Dataset Results. Table 2 shows TOLA leading all six metrics, exceeding DiffTSR by 1.5953 dB PSNR, 4.50 points TransOCR ACC, and 8.00 points PaddleOCR ACC.

Evaluation on RT50. Figure 4 shows that TOLA better preserves strokes and character shapes on Chinese, English, and numeric text under varied real degradations.

Inference Cost and Latency. We use synchronized batch-1 timing on 100 fixed CTR images with one NVIDIA RTX PRO 6000, excluding loading. Table 3 reports complete-trajectory MACs at 128×512 and restoration-module parameters excluding VAEs and auxiliary text modules. TOLA uses 297.497 G MACs and 0.1319 seconds per image, versus 59,213.970 G and 9.9548 seconds for DiffTSR, while its optimized and saved state is 2.46M parameters. Figure 5 visualizes the quality-latency trade-off.

Ablations and Diagnostic Analyses

Core Components. Table 4 shows that LoRA drives one-step adaptation and LRC adds 0.3847 dB PSNR while reducing LPIPS/DISTS by 0.0121/0.0017; its ACC change is not significant ($p = 0.584$). A parameter-matched plain correction loses 0.8249 dB PSNR and increases FID by 4.2491, so capacity alone does not explain the gain. IDM-only LoRA remains much closer to Full than VAE-only LoRA.



Figure 3: Comparison on five selected CTR-TSR-Test $\times 4$ images covering Chinese, English, and numeric text. Each column presents an aligned test sample, and the rows show the LR input, displayed baselines, Ours, and the HR reference.

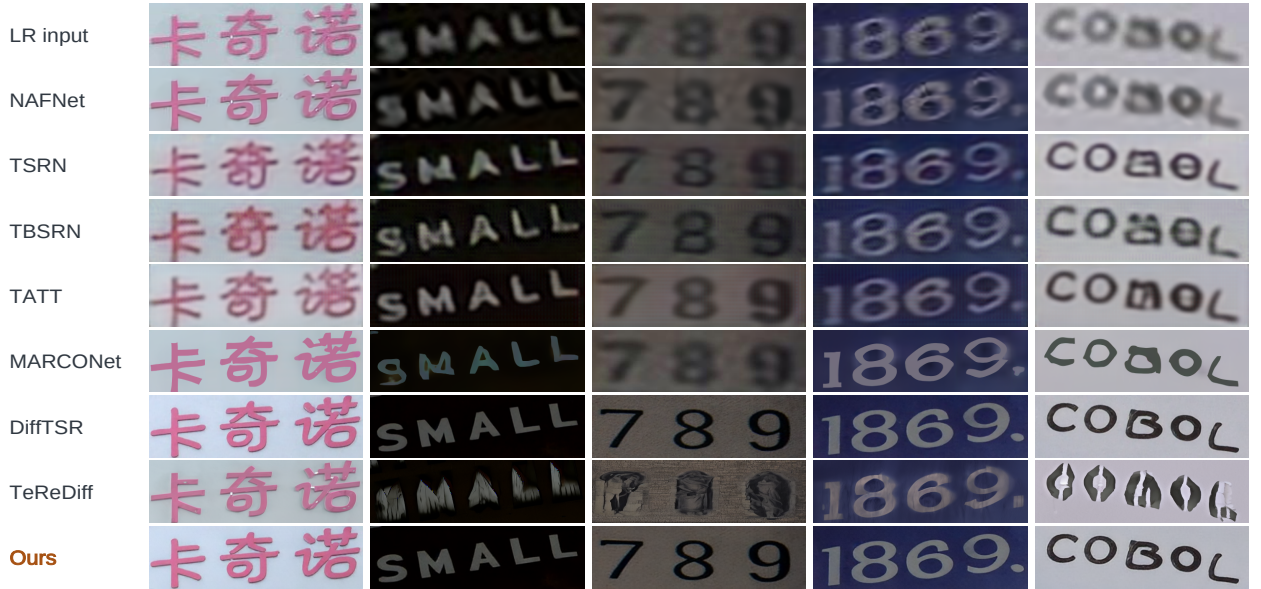


Figure 4: RT50 comparison on Chinese, English, and numeric text under varied real degradations. Rows show LR inputs, displayed baselines, and Ours; examples cover blur, illumination, viewpoint, resolution, and background variations.

Removing MoM or bypassing its latent fusion degrades every metric. Relative to uniform confidence, confidence weighting adds 0.1908 dB PSNR and 0.0134/0.0138 TransOCR ACC/NED while reducing LPIPS/DISTS/FID by 0.0213/0.0160/4.2766. Null context confirms the value of predicted text, and the GT oracle adds another 0.0862 ACC and 0.0716 NED, revealing headroom from better priors.

Objective and Hyperparameter Diagnostics. Rank 4 outperforms ranks 2 and 8. Small, default, and

large LRCs use (hidden, growth, RRDBs) = (16, 8, 1), (32, 16, 1), and (32, 16, 2) with 0.045235M, 0.180323M, and 0.360323M parameters; the default gives the best trade-off. Removing ℓ_1 most reduces PSNR, removing LPIPS most harms perceptual metrics, and OCR weight 0 lowers recognition; 0.05 brings no gain. Among tested timesteps, $t = 999$ performs best.

Text-Prior Reliability and Error Analysis. Figure 7 shows TOLA outperforms DiffTSR across confidence bins and text categories, reducing substitutions, deletions,

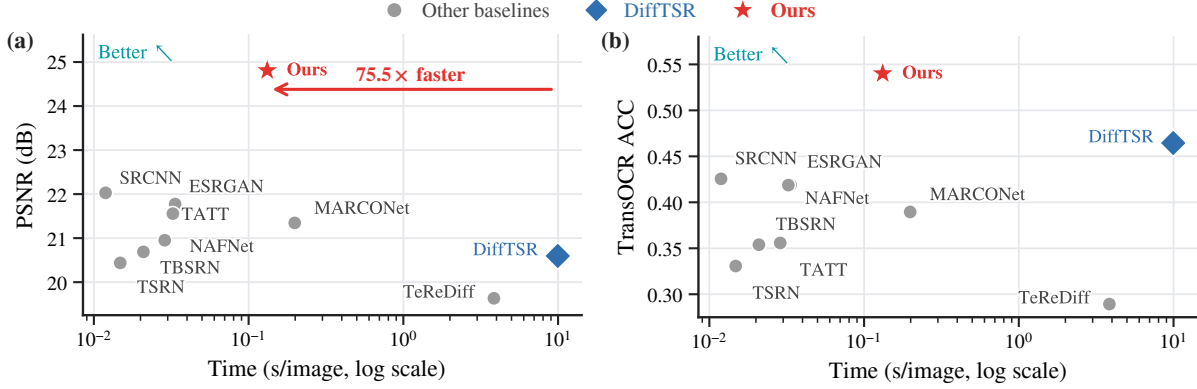


Figure 5: Quality–latency comparison on CTR-TSR-Test $\times 4$. Panels (a) and (b) plot PSNR and TransOCR ACC against synchronized batch-1 inference time measured on 100 fixed CTR inputs using one NVIDIA RTX PRO 6000.

(a) Complementary roles of LoRA and LRRB



(b) Text-conditioning pathway



Figure 6: Core component visualization. (a) LoRA enables one-step adaptation and LRRB corrects local structure over frozen and single-module controls. (b) No MoM and direct-token conditioning cause character errors; LR and HR provide references.

Method	# Params. (M)	MACs (G)	Inference Time (s)
SRCNN	0.057	3.748	0.0119
ESRGAN	16.70	73.453	0.0335
NAFNet	67.89	63.195	0.0324
TSRN	2.68	0.904	0.0148
TBSRN	3.21	2.535	0.0209
TATT	7.61	1.269	0.0287
MARCONet	87.90	466.178	0.1986
DiffTSR	874.00	59,213.970	9.9548
TeReDiff	1682.54	26,839.909	3.8499
Ours	876.23	297.497	0.1319

Table 3: Restoration parameters, MACs, and batch-1 inference time on 100 CTR images. VAEs and text modules are excluded; TOLA stores 2.46M adapted parameters.

and insertions. The exact-match gap between correct and incorrect priors identifies recognition errors as the main limitation.

Inference Controls and Stochastic Stability. Controls favor $t = 999$; zero noise and disabled VAE LoRA reduce quality, while LR-input and frozen-VAE controls rule out copying or VAE-only reconstruction. Across three draws, maximum standard deviations are 0.0019 dB PSNR and 0.0012 ACC.

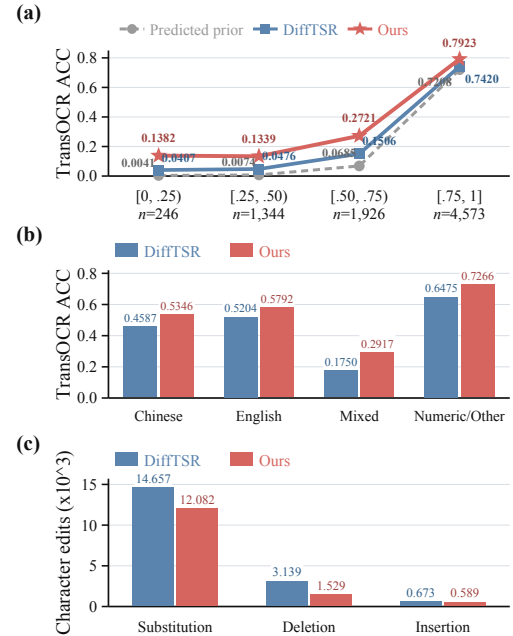


Figure 7: Recognition diagnostics on CTR-TSR-Test $\times 4$: (a) ACC across confidence bins, (b) ACC across text categories, and (c) character-level edit counts.

Configuration	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FID $\downarrow$	T-ACC $\uparrow$	T-NED $\uparrow$	P-ACC $\uparrow$	P-NED $\uparrow$
Frozen one-step	20.0416	0.5342	0.5254	0.3464	119.4517	0.2276	0.3925	0.1278	0.2918
LRC only	21.9696	0.6021	0.4880	0.3301	110.0081	0.3785	0.5713	0.2440	0.4337
LoRA only	24.4220	0.7109	0.2410	0.1970	12.7420	0.5378	0.7337	0.4570	0.6564
Full	24.8095	0.7235	0.2288	0.1953	12.6849	0.5401	0.7370	0.4626	0.6585
Plain residual correction	23.9846	0.6991	0.2843	0.2180	16.9340	0.5080	0.7047	0.4256	0.6243
No MoM	23.9624	0.6984	0.2833	0.2170	16.1812	0.5060	0.7049	0.4117	0.6197
Direct token condition	24.0026	0.6993	0.2846	0.2186	16.7650	0.5079	0.7038	0.4093	0.6185
IDM LoRA only	23.8812	0.6987	0.2844	0.2174	16.7593	0.5104	0.7058	0.4232	0.6254
VAE LoRA only	22.4861	0.6141	0.4635	0.3221	108.0600	0.3989	0.5926	0.2597	0.4505
Null context	24.5495	0.7118	0.2695	0.2239	22.1417	0.5308	0.7280	0.4390	0.6438
Uniform confidence	24.6159	0.7149	0.2502	0.2113	16.9615	0.5258	0.7224	0.4285	0.6356
GT prior (oracle)	24.9148	0.7283	0.2216	0.1920	12.7565	0.6253	0.8079	0.5520	0.7387

Table 4: Ablation on CTR-TSR-Test $\times 4$. Full uses IDM/VAE LoRA, LRC, and confidence-aware MoM. Plain is parameter-matched; No MoM removes the text branch; direct tokens bypass fusion; GT is oracle; T/P denote TransOCR/PaddleOCR.

Paired Uncertainty. Paired bootstrap intervals exclude zero for LRC effects on PSNR, SSIM, LPIPS, and DISTS, but not NED; exact McNemar testing finds no significant ACC change. Predicted conditioning improves ACC over null context, and the GT-prior gap confirms prior errors.

Conclusion

We presented TOLA, a text-aware one-step adaptation of DiffTSR. TOLA constructs a confidence-aware TransOCR–MoM condition once, adapts IDM and the VAE encoder with rank-4 LoRA, and corrects clean-latent residual errors with LRRB without invoking TDM. It improves all nine CTR-TSR-Test metrics over DiffTSR while reducing IDM/TDM evaluations from 200/200 to 1/0 and latency from 9.9548 to 0.1319 seconds per image. The optimized and saved state contains only 2.46M parameters, while inference retains the pretrained backbone without introducing another iterative restoration trajectory. Component ablations and independent-OCR evaluation support the complementary roles of LoRA, LRRB, and confidence-aware conditioning. Results on RealCE-200 and RT50 further demonstrate generalization across paired camera observations, Chinese, English, and numeric text, and varied real capture conditions. These results establish efficient one-step diffusion TSR without iterative image–text refinement.

References

Jingye Chen, Bin Li, and Xiangyang Xue. Scene Text Telescope: Text-focused scene image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12026–12035, 2021.

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, pages 285–293, 2022a.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European Conference on Computer Vision*, pages 17–33, 2022b.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022.

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, 2016.

Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. TSD-SR: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23174–23184, 2025.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. PP-OCR: A practical ultra lightweight OCR system, 2020.

Hang Guo, Tao Dai, Guanghao Meng, and Shu-Tao Xia. Towards robust scene text image super-resolution via explicit location enhancement. In *International Joint Conference on Artificial Intelligence*, pages 782–790, 2023.

- Haodong He, Xin Zhan, Yancheng Bai, Rui Lan, Lei Sun, and Xiangxiang Chu. TEXTS-Diff: TEXTS-aware diffusion model for real-world text image super-resolution. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4306–4310, 2026.
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pages 6626–6637, 2017.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.
- Qiming Hu, Linlong Fan, Yiyan Luo, Yuhang Yu, Xiaojie Guo, and Qingnan Fan. Text-aware real-world image super-resolution via diffusion model with joint segmentation decoders. In *Advances in Neural Information Processing Systems*, volume 38, pages 61522–61543, 2025.
- Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, volume 35, pages 23593–23606, 2022.
- Aro Kim, Myeongjin Jang, Chaewon Moon, Youngjin Shin, Jinwoo Jeong, and Sang-hyo Park. FiDeSR: High-fidelity and detail-preserving one-step diffusion super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 38270–38280, 2026.
- Yuxin Kong, Weihong Ma, Lianwen Jin, and Yang Xue. GARDEN: Generative prior guided network for scene text image super-resolution. In *International Conference on Document Analysis and Recognition*, volume 14808, pages 196–214, 2024.
- 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*, pages 10103–10113, 2023.
- Jingyun Liang, Jiezhong 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 Workshops*, pages 1833–1844, 2021.
- Baolin Liu, Zongyuan Yang, Chinwai Chiu, and Yongping Xiong. TextDiff: Enhancing scene text image super-resolution with mask-guided residual diffusion models. *Pattern Recognition*, 164:111513, 2025.
- Jianqi Ma, Zhetong Liang, and Lei Zhang. 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*, pages 5911–5920, 2022.
- Jianqi Ma, Shi Guo, and Lei Zhang. Text prior guided scene text image super-resolution. *IEEE Transactions on Image Processing*, 32:1341–1353, 2023a.
- Jianqi Ma, Zhetong Liang, Wangmeng Xiang, Xi Yang, and Lei Zhang. A benchmark for chinese-english scene text image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19452–19461, 2023b.
- Jaewon Min, Jin Hyeon Kim, Paul Hyunbin Cho, Jaeun Lee, Jihye Park, Minkyu Park, Sangpil Kim, Hyunhee Park, and Seungryong Kim. Text-aware image restoration with diffusion models. In *International Conference on Learning Representations*, 2026.
- Axi Niu, Kang Zhang, Qingsen Yan, Hao Jin, Jinqiu Sun, and Yanning Zhang. DualTSR: Unified dual-diffusion transformer for scene text image super-resolution, 2026.
- 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*, pages 1485–1495, 2024.
- Chenglu Pan, Xiaogang Xu, Ganggui Ding, Yunke Zhang, Wenbo Li, Jiarong Xu, and Qingbiao Wu. Boosting diffusion-based text image super-resolution model towards generalized real-world scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 7459–7468, 2025.
- Dongwoo Park and Suk Pil Ko. NCAP: Scene text image super-resolution with Non-Categorical prior. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2432–2441, 2025.

- 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*, pages 10684–10695, 2022.
- Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2023.
- Shrey Singh, Prateek Keserwani, Masakazu Iwamura, and Partha Pratim Roy. DCDM: Diffusion-conditioned-diffusion model for scene text image super-resolution. In *European Conference on Computer Vision*, pages 303–320, 2024.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.
- Ying Tai, Rui Xie, Chen Zhao, Kai Zhang, Zhenyu Zhang, Jun Zhou, and Jian Yang. AddSR: Accelerating diffusion-based blind super-resolution with adversarial diffusion distillation. *Pattern Recognition*, 175:113012, 2026.
- LeoWu TomyEnrique, Xiangcheng Du, Kangliang Liu, Han Yuan, Zhao Zhou, and Cheng Jin. Efficient scene text image super-resolution with semantic guidance. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3160–3164, 2024.
- Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024a.
- Wenjia Wang, Enze Xie, Xuebo Liu, Wenhai Wang, Ding Liang, Chunhua Shen, and Xiang Bai. Scene text image super-resolution in the wild. In *European Conference on Computer Vision*, pages 650–666, 2020.
- Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. ESRGAN: Enhanced super-resolution generative adversarial networks. In *European Conference on Computer Vision Workshops*, pages 63–79, 2018.
- 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 Workshops*, pages 1905–1914, 2021.
- Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. SinSR: Diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25796–25805, 2024b.
- Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- Baole Wei, Yuxuan Zhou, Liangcai Gao, and Zhi Tang. GlyphSR: A simple glyph-aware framework for scene text image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8277–8285, 2025.
- Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. In *Advances in Neural Information Processing Systems*, volume 37, pages 92529–92553, 2024.
- Zihang Xu, Xiaoyang Liu, Zheng Chen, Yulun Zhang, and Xiaokang Yang. PRISM: Prior rectification and uncertainty-aware structure modeling for diffusion-based text image super-resolution, 2026.
- Wei Yang, Yihong Luo, Mayire Ibrayim, and Askar Hamdulla. More and less: Enhancing abundance and refining redundancy for text-prior-guided scene text image super-resolution. In *International Conference on Document Analysis and Recognition*, volume 14808, pages 129–146, 2024.
- Keren Ye, Ignacio Garcia Dorado, Michalis Raptis, Mauricio Delbracio, Irene Zhu, Peyman Milanfar, and Hossein Talebi. TextSR: Diffusion super-resolution with multilingual OCR guidance, 2025.
- Shengrong Yuan, Runmin Wang, Ke Hao, Xuqi Ma, Changxin Gao, Li Liu, and Nong Sang. StyleSRN: Scene text image super-resolution with text style embedding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18693–18702, 2025.
- Zongsheng Yue, Jianyi Wang, and Chen Change Loy. ResShift: Efficient diffusion model for image super-resolution by residual shifting. In *Advances in Neural Information Processing Systems*, volume 36, pages 13294–13307, 2023.
- Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the*

- IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.
- Yuzhe Zhang, Jiawei Zhang, Hao Li, Zhouxia Wang, Luwei Hou, Dongqing Zou, and Liheng Bian. Diffusion-based blind text image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25827–25836, 2024.
- Cairong Zhao, Shuyang Feng, Brian Nlong Zhao, Zhijun Ding, Jun Wu, Fumin Shen, and Heng Tao Shen. Scene text image super-resolution via parallelly contextual attention network. In *ACM International Conference on Multimedia*, pages 2908–2917, 2021.
- Minyi Zhao, Miao Wang, Fan Bai, Bingjia Li, Jie Wang, and Shuigeng Zhou. C3-STISR: Scene text image super-resolution with triple clues. In *International Joint Conference on Artificial Intelligence*, pages 1707–1713, 2022.
- Yuxuan Zhou, Liangcai Gao, Zhi Tang, and Baole Wei. Recognition-guided diffusion model for scene text image super-resolution. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2940–2944, 2024.
- Shipeng Zhu, Zuoyan Zhao, Pengfei Fang, and Hui Xue. Improving scene text image super-resolution via dual prior modulation network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3843–3851, 2023a.
- Xiangyuan Zhu, Kehua Guo, Hui Fang, Rui Ding, Zheng Wu, and Gerald Schaefer. Gradient-based graph attention for scene text image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3861–3869, 2023b.

Appendices

This supplement provides implementation and evaluation details complementary to the self-contained main paper. It clarifies the system boundary relative to nearby diffusion methods, enumerates the modules executed by the one-step route, specifies the saved adaptation state, datasets, metrics, and controlled variants, and reports the exact values underlying the extended diagnostics. It also presents additional ablations, inference controls, uncertainty analysis, and qualitative comparisons.

A System and Comparison Boundaries

A.1 One-Step Inference Route

For an input image, the VAE encoder first produces the unscaled three-channel latent z_{LR} . TOLA uses the zero-indexed timestep $t = 999$ of the 1000-step linear schedule inherited from DiffTSR, whose endpoints are $\beta_0 = 0.0015$ and $\beta_{999} = 0.0205$. One Gaussian sample is drawn to construct

$$z_t = \sqrt{\alpha_t} z_{\text{LR}} + \sqrt{1 - \alpha_t} \epsilon.$$

TransOCR is evaluated once. After Simplified-Chinese normalization, its token embeddings are multiplied by the aligned raw token confidences and passed to the frozen MoM together with $0.18215[z_{\text{LR}}, z_t]$. MoM computes both inherited image and text conditions once, but the one-step IDM uses only the text condition as cross-attention context.

IDM receives $[z_t, z_{\text{LR}}]$ and predicts noise once. Because the official DiffTSR checkpoint uses epsilon prediction, the initial clean latent is reconstructed as

$$\hat{z}_0 = \frac{z_t - \sqrt{1 - \alpha_t} \hat{\epsilon}}{\sqrt{\alpha_t}}.$$

There is no reverse chain after this conversion. Latent residual correction (LRC) uses a compact correction branch to correct the LR-to-clean residual $r = z_{\text{LR}} - \hat{z}_0$:

$$\Delta r = g_\phi([z_{\text{LR}}, r]), \quad \hat{z}_{0, \text{corr}} = z_{\text{LR}} - (r + \Delta r).$$

Finally, the frozen decoder maps $\hat{z}_{0, \text{corr}}/0.18215$ to x_{SR} . Each image is restored once, without multi-sample selection or averaging.

B Architecture and Saved State

B.1 LoRA Placement and Frozen Components

Rank-4 LoRA adapters use $\alpha = 4$ and zero dropout. They are inserted into eligible linear and convolutional layers of IDM and the VAE encoder, including its encoder-side 1×1 latent projection. The first input convolution of IDM, the decoder-side latent projection, and the complete VAE decoder remain unadapted. TransOCR, MoM, all base IDM/VAE weights, and the loaded but unused TDM decoder remain frozen. The conditioning TransOCR and the OCR-loss recognizer are separate frozen instances initialized from the same checkpoint.

Trainable partition	Parameters	Saved?
IDM LoRA	2,027,520	yes
VAE-encoder LoRA	250,548	yes
LRC	180,323	yes
Total adaptation state	2,458,391	yes

Table S3: Optimized and saved parameters.

B.2 Latent Residual Correction

TOLA implements LRC with a compact correction branch inspired by FiDeSR. It receives the six-channel concatenation $[z_{\text{LR}}, r]$. A 1×1 input projection maps 6 channels to 32, followed by one residual-in-residual dense block (RRDB) and a 1×1 output projection from 32 channels to 3. The RRDB contains three residual dense blocks; each residual dense block contains five densely connected 3×3 convolutions with growth width 16. Residual scaling is 0.2 at both dense-block and RRDB levels. The output projection is zero initialized, so LRC initially leaves $\hat{z}_0$ unchanged.

Small, Default, and Large LRC use $(h, g, n_{\text{RRDB}}) = (16, 8, 1)$, $(32, 16, 1)$, and $(32, 16, 2)$, respectively, with 0.045235M, 0.180323M, and 0.360323M parameters. The parameter-matched baseline uses a plain correction branch with a $6 \rightarrow 30$ projection, 11 two-convolution residual blocks with LeakyReLU and residual scaling 0.2, a $30 \rightarrow 30$ mixing layer, and a zero-initialized $30 \rightarrow 3$ projection (0.180093M parameters).

C Training Protocol

Computing Environment. Experiments run on Ubuntu 22.04 with an AMD EPYC 9J14 96-Core Processor, 503 GiB of system memory, and NVIDIA RTX PRO 6000 Blackwell Server Edition GPUs with 97,887 MiB per GPU. The NVIDIA driver version is 580.142. The primary environment uses Python 3.10.12, PyTorch 2.11.0+cu128, Torchvision 0.26.0+cu128, CUDA 12.8, cuDNN 9.19, PyIQA 0.1.15.post2, and LPIPS 0.1.4. The independent OCR evaluation uses PaddlePaddle 2.6.2 and PaddleOCR 2.7.3. These package versions define the software environment used for the reported results.

Training reads 63,644 high-quality CTR preprocessing images and applies a deterministic local 98%/2% partition: 62,371 images for training and 1,273 for validation. This is a local partition rather than an official CTR split. Each image is degraded online with equal probability by the Real-ESRGAN or BSRGAN pipeline (Wang et al., 2021; Zhang et al., 2021); the scale is drawn from $\{1, 2, 4\}$. Both degraded inputs and targets use a 128×512 canvas.

The complete objective is

$$\mathcal{L} = \mathcal{L}_1(x_{\text{SR}}, x_{\text{HR}}) + \mathcal{L}_{\text{LPIPS-VGG}}(x_{\text{SR}}, x_{\text{HR}}) + 0.02 \mathcal{L}_{\text{OCR}}.$$

The frozen OCR recognizer remains differentiable with respect to its image input, so OCR-loss gradients reach LoRA and LRC. No latent ℓ_1 , noise-prediction, score-distillation, VSD/CSD, or teacher-regularization loss is used.

Method	Image route	Text route	Distinction
DiffTSR	Iterative IDM	TDM–MoM iterations	Coupled image/text diffusion
Boosting DiffTSR	Iterative diffusion	Confidence-guided TDM–MoM	SR prior + progressive sampling
OSDiff	One-step latent	Degradation prompt	LoRA + VSD
FiDeSR	One-step residual	No OCR	LRC + detail/frequency modules
PRISM	Single-pass flow	Rectified prior	Flow matching + structure modeling
TADiSR	Text-aware diffusion	Text segmentation	Joint segmentation decoders
TEXTS-Diff	One-step latent	Abstract + region cues	Real-Texts + text-aware restoration
DualTSR	Conditional flow	Discrete diffusion	Shared multimodal transformer
TOLA (Ours)	One IDM call	TransOCR–MoM once	One-Step; no TDM

Table S1: Mechanistic positioning of the closest diffusion routes. Methods follow (Zhang et al., 2024; Pan et al., 2025; Wu et al., 2024; Kim et al., 2026; Xu et al., 2026; Hu et al., 2025; He et al., 2026; Niu et al., 2026). The table distinguishes system boundaries.

Component	Function	Trainable update	Calls/image
TransOCR	token and confidence prediction	frozen	1
VAE encoder	LR latent encoding	LoRA	1
MoM	text-aware condition construction	frozen	1
IDM U-Net	one-step diffusion prediction	LoRA	1
LRC	latent residual correction	full branch	1
VAE decoder	corrected-latent decoding	frozen	1

Table S2: Inference boundary of TOLA. TransOCR, MoM, and IDM are each executed once; TDM is not executed.

Item	Setting
Initialization	official DiffTSR CTR checkpoint
Training steps	100,000
GPUs	4× RTX PRO 6000
Batch/GPU	16
Effective batch	64
Precision	FP32
Training time	approximately 55 hours
Optimizer	AdamW
Learning rate	5×10^{-5}
Adam betas	(0.9, 0.999)
Weight decay	0.01
Warm-up	500 linear steps
Gradient clipping	1.0
Selection	validation performance

Table S4: Primary training configuration.

C.1 Hyperparameter Search and Selection

Table S5 summarizes the search. The final architecture, objective, OCR-loss weight, and training timestep are selected using validation performance. Most candidate checkpoints use validation LPIPS; the no-LPIPS objective uses validation PSNR. The remaining optimization settings are reported in Table S4.

D Dataset Appendix and Evaluation Metrics

D.1 Dataset Inventory

Table S6 summarizes the evaluation sets. The primary benchmark is the complete 8,089-image CTR-TSR-Test ×4

set. All methods under the unified protocol are filename-matched to the same references and evaluated on a 128×512 output canvas.

D.2 Motivation and Coverage

Existing benchmarks do not individually cover all conditions required for evaluating real scene-text restoration. CTR-TSR-Test provides a large paired benchmark but primarily reflects synthetic degradations, whereas available real-image datasets differ in language coverage, capture conditions, annotation format, and reference availability. RealCE-200 provides a paired protocol constructed from real multi-focal observations, while RT50 balances Chinese, English, and numeric text collected through heterogeneous acquisition routes. Together, these sets broaden the evaluated conditions.

D.3 RealCE-200 Construction

We construct RealCE-200 from the multi-focal images and official text annotations in the RealCE validation split before evaluation. For each official text annotation, identical bounding-box coordinates are applied to the corresponding 13 mm and 52 mm focal images, which serve as LR and reference observations. Both crops are bicubically resized to 128×512 without geometric registration. Automatic filtering considers transcription length and composition, bounding-box size, aspect ratio, sharpness, and contrast. Duplicate candidates are removed using transcription identity and perceptual dHash. Official transcriptions are retained after whitespace cleaning, and no manual relabeling is performed. The resulting records use fixed identifiers R001–R200. RealCE is not used for training, validation, or checkpoint selection.

Hyperparameter	Values examined	Count	Final setting
LoRA (r, α)	(2, 2), (4, 4), (8, 8)	3	(4, 4)
LRC (h, g, n_{RRDB})	(16, 8, 1), (32, 16, 1), (32, 16, 2)	3	(32, 16, 1)
Reconstruction objective	Full, without ℓ_1 , without LPIPS	3	Full
OCR-loss weight	{0, 0.02, 0.05}	3	0.02
Training timestep	{899, 949, 999}	3	999

Table S5: Hyperparameter and objective settings examined during development and the final configuration used by TOLA.

Set	Source	Size	Content and reference
CTR-TSR-Test $\times 4$	Official CTR-TSR-Test	8,089 pairs	Filename-aligned LR/HR images and transcriptions
RealCE-200	RealCE multi-focal images (Ma et al., 2023b)	200 pairs	Paired 13 mm LR/52 mm reference crops
RT50	Multi-source real images	50 inputs	17 Chinese, 17 English, and 16 numeric samples

Table S6: Evaluation datasets. RealCE-200 and RT50 are constructed in this work.

D.4 RT50 Construction

RT50 contains 50 real text inputs collected through multiple acquisition routes. The fixed composition contains 17 Chinese samples, 17 English samples, and 16 numeric samples; the difficulty partition contains 10 mild, 20 medium, and 20 hard samples. Candidate labels are restricted to 2–8 Chinese characters, 3–12 English letters, or 2–6 digits. Images must be at least 24×8 pixels with aspect ratios between 2 and 10. After canonical resizing, admissible candidates have gray mean in $[20, 235]$, gray standard deviation in $[10, 80]$, Laplacian variance in $[2, 1200]$, edge density in $[0.01, 0.65]$, and entropy of at least 3.0. Repeated transcriptions are removed, and near duplicates are rejected at a dHash Hamming-distance threshold of 7. The selection protocol is finalized before model inference without consulting restoration or OCR outputs. Records use fixed identifiers RT001–RT050 and are normalized to the same 128×512 model-input canvas.

D.5 Included Dataset Records

The public data release will include the exact image files evaluated in this work rather than requiring users to reconstruct them. For RealCE-200, it will provide all 200 LR/reference pairs together with the pair manifest, transcriptions, source filenames, and SHA-256 values. For RT50, it will provide all 50 input images together with the complete manifest, labels, provenance table, selection protocol, per-image geometry and quality statistics, category and difficulty assignment, dHash, and SHA-256 values. The accompanying records define the evaluated samples independently of model outputs and verify every released file. Upon publication, these exact datasets and their metadata will be publicly available under research-use terms documented with the release.

D.6 Metric Implementations

Predictions and references are matched by exact filename. PyIQA computes PSNR and SSIM on luminance without

border crop. Evaluation LPIPS uses the AlexNet backbone, while training LPIPS uses VGG. DISTS and FID use PyIQA; FID is computed once over the complete output and reference directories.

The primary recognizer is TransOCR. Predictions undergo full-width to half-width conversion, Simplified-Chinese conversion, and whitespace removal; case is preserved. Exact-match accuracy is denoted ACC. Per-image normalized edit-distance similarity is

$$\text{NED} = 1 - \frac{\text{ED}(p, g)}{\max(|p|, |g|, 1)},$$

and is averaged across the set. The independent cross-check uses PaddleOCR 2.7.3 with PaddlePaddle 2.6.2 and the Chinese PP-OCRv4 recognition model; detection and orientation classification are disabled. This recognizer is not used for conditioning, training, validation, or checkpoint selection.

E Extended Qualitative Comparisons

E.1 Additional RealCE Comparison

Figure S1 provides additional paired RealCE-200 examples under the unified evaluation protocol.

E.2 BTL Cross-Protocol Comparison

Figure S2 compares TOLA, DiffTSR, TeReDiff, and PRISM on five BTL-test $\times 4$ examples (Xu et al., 2026). The examples cover Chinese text, English words, and numeric sequences.

F Controlled Variant Definitions

All filename-paired comparisons use identical per-image Gaussian noise for the two compared outputs. Effect-size intervals use 5,000 filename-paired bootstrap resamples. Exact-match ACC is additionally tested with the exact paired McNemar test. Table S7 defines all architecture and conditioning controls.

LR input	环保投诉举报电话		地址：青松路125号	13958832240	Blockchain
SRCNN	环保投诉举报电话		地址：青松路125号	13958832240	Blockchain
ESRGAN	环保投诉举报电话		地址：青松路125号	13958832240	Blockchain
NAFNet	环保投诉举报电话		地址：青松路125号	13958832240	Blockchain
TSRN	环保投诉举报电话		地址：青松路125号	13958832240	Blockchain
TBSRN	环保投诉举报电话		地址：青松路125号	13958832240	Blockchain
TATT	环保投诉举报电话		地址：青松路125号	13958832240	Blockchain
MARCONet	环保投诉举报电话		地址：青松路125号	13958832240	Blockchain
DiffTSR	环保投诉举报电话		地址：青松路125号	13958832240	Blockchain
TeReDiff	环保投诉举报电话		地址：青松路125号	13958832240	Blockchain
Ours	环保投诉举报电话		地址：青松路125号	13958832240	Blockchain
HR	环保投诉举报电话		地址：青松路125号	13958832240	Blockchain

Figure S1: Qualitative comparison on five paired RealCE-200 examples. Rows show the LR input, SRCNN, ESRGAN, NAFNet, TSRN, TBSRN, TATT, MARCONet, DiffTSR, TeReDiff, Ours, and the HR reference.

LR input					
DiffTSR					
TeReDiff					
PRISM					
Ours					
HR					

Figure S2: Cross-protocol comparison on five BTL-test $\times 4$ examples. Rows show the LR input, DiffTSR, TeReDiff (Min et al., 2026), PRISM (Xu et al., 2026), Ours, and GT. BTL-test combines real and rendered HQ sources with BSRGAN/Real-ESRGAN degradations; the PRISM row is reproduced from the original paper.

Variant	Definition
Frozen one-step	Frozen DiffTSR pathway with one IDM call and no LoRA or LRC.
LRC only	Default LRC is optimized while the pretrained image pathway remains frozen.
LoRA only	Rank-4 IDM and VAE-encoder LoRA without LRC.
Full	Rank-4 LoRA plus the default LRC and confidence-aware MoM condition.
No MoM	The TransOCR–MoM conditioning branch is removed during training and inference.
Direct tokens	Predicted token embeddings are supplied without MoM latent fusion.
Null context	Fixed-checkpoint control retaining MoM with a null text context.
Uniform confidence	Fixed-checkpoint control retaining predicted tokens and MoM with unit confidences.
GT prior	Fixed-checkpoint oracle replacing predicted tokens with ground-truth text.
IDM LoRA only	VAE-encoder LoRA is omitted; IDM LoRA and default LRC remain.
VAE LoRA only	IDM LoRA is omitted; VAE-encoder LoRA and default LRC remain.
VAE LoRA off	Inference disables all VAE-encoder LoRA, including the latent projection, while retaining IDM LoRA and LRC.

Table S7: Architecture and conditioning controls used in the diagnostics.

Configuration	PSNR	SSIM	LPIPS	DISTS	FID
Full	24.8095	.7235	.2288	.1953	12.6849
Rank 2	23.9287	.6979	.2835	.2168	16.8514
Rank 8	23.9573	.6993	.2815	.2179	17.2784
LRC Small	23.9506	.6984	.2870	.2192	17.4977
LRC Large	23.9967	.6990	.2781	.2154	15.6799
No ℓ_1	23.1142	.6906	.2801	.2153	15.6182
No LPIPS	24.0026	.6888	.3820	.2744	50.5338
OCR wt. 0	23.9157	.6997	.2741	.2132	16.0079
OCR wt. .05	23.9290	.6961	.2955	.2232	18.1809
Train $t = 899$	24.2532	.7088	.2571	.2101	13.5299
Train $t = 949$	24.1818	.7068	.2644	.2121	13.8379

Table S8: Image-quality diagnostics on CTR-TSR-Test $\times 4$. The reference uses rank 4, the default 0.180M LRC, OCR weight 0.02, and $t = 999$.

Inference Runs. The main comparisons, ablations, and controls use one restored output per image without multi-sample selection or output averaging. The stochastic-stability analysis contains three inference runs and reports their mean and standard deviation. Bootstrap resampling does not generate additional restored outputs.

G Extended Ablation and Control Results

G.1 Objective and Hyperparameter Diagnostics

Tables S8 and S9 show that rank 4 provides a better overall balance than ranks 2 and 8. The default LRC also outperforms its smaller and larger variants overall, indicating that additional capacity is not the source of the gain. Removing ℓ_1 causes the largest PSNR reduction, whereas removing LPIPS most strongly degrades perceptual and distributional metrics. An OCR weight of 0.02 outperforms both zero weight and 0.05, and $t = 999$ gives the strongest overall training result among the tested timesteps.

G.2 Prior-Stratified Performance

Table S10 separates samples according to whether the predicted TransOCR prior matches the ground truth.

Statistic	Correct prior	Wrong prior
N	3,439	4,650
Mean confidence	0.9536	0.6015
DiffTSR ACC	0.9514	0.1043
TOLA ACC	0.9683	0.2217
DiffTSR NED	0.9872	0.4038
TOLA NED	0.9906	0.5482
TOLA PSNR	26.2307	23.7535
TOLA LPIPS	0.1652	0.2760

Table S10: Performance stratified by TransOCR prior correctness, with image and recognition metrics.

G.3 Inference Controls and Stochastic Stability

Table S11 shows that $t = 949$ slightly increases PSNR and SSIM but worsens LPIPS, DISTS, and FID, while $t = 899$ further degrades perceptual and distributional quality. The

Configuration	T-ACC	T-NED	P-ACC	P-NED
Full	.5401	.7370	.4626	.6585
Rank 2	.5138	.7056	.4253	.6247
Rank 8	.5119	.7052	.4211	.6224
LRC Small	.5079	.7038	.4235	.6235
LRC Large	.5097	.7054	.4247	.6240
No ℓ_1	.5092	.7031	.4248	.6237
No LPIPS	.5044	.6987	.4015	.6049
OCR wt. 0	.5025	.6981	.4253	.6216
OCR wt. .05	.5096	.7042	.4217	.6234
Train $t = 899$	.5127	.7063	.4249	.6266
Train $t = 949$	.5107	.7060	.4271	.6256

Table S9: Recognition diagnostics for the same configurations. Most checkpoints use validation LPIPS; the no-LPIPS variant uses validation PSNR.

default $t = 999$ therefore provides the best overall balance. Zero noise and disabled VAE LoRA substantially reduce reconstruction quality, and the input controls rule out copying or a frozen-VAE round trip as explanations for the gains. Repeated inference shows low sensitivity to the sampled Gaussian noise.

Efficiency Measurement Details. All methods are timed on the same 100 CTR-TSR-Test images at batch size 1 on one NVIDIA RTX PRO 6000 GPU. Timing follows warm-up, excludes loading, and synchronizes CUDA. TOLA timing includes VAE encoding, TransOCR, MoM, one IDM evaluation, LRC, and VAE decoding; DiffTSR and TeReDiff use their full 200- and 50-step routes. MACs sum restoration-backbone evaluations at 128×512 . Counts exclude VAEs and auxiliary recognition, detection, and spotting modules, so the reported 876.23M restoration-module count differs from the 2.46M optimized and saved adaptation state.

Planned Public Release. Upon publication, we will release the implementation, configurations, and exact RealCE-200 and RT50 image records under research-use terms. The release will document the required third-party checkpoints without redistributing their pretrained weights.

G.4 Paired Uncertainty

Table S12 reports paired intervals, which exclude zero for the LRC effects on PSNR, SSIM, LPIPS, and DISTS, but not NED; the exact paired McNemar test likewise finds no significant ACC difference. Because all comparisons use filename-aligned outputs, the intervals measure within-sample changes and separate consistent reconstruction gains from limited recognition changes. Paired resampling preserves image correspondence when estimating uncertainty across methods. Predicted conditioning significantly improves ACC over null context, whereas the GT-prior gap confirms remaining headroom from prior errors. Figure S3 plots these conditioning controls, and Figure S4 visualizes the paired LRC effects.

(a) Fixed-checkpoint image metrics						(b) Fixed-checkpoint recognition metrics				
Control	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	DISTS $\downarrow$	FID $\downarrow$	Control	T-ACC $\uparrow$	T-NED $\uparrow$	P-ACC $\uparrow$	P-NED $\uparrow$
Inference $t = 999$	24.8067	.7235	.2289	.1953	12.6849	Inference $t = 999$	.5391	.7362	.4626	.6585
Inference $t = 949$	24.9307	.7257	.2536	.2109	17.8904	Inference $t = 949$	.5402	.7381	.4614	.6589
Inference $t = 899$	24.6970	.7219	.2777	.2250	22.6079	Inference $t = 899$	.5395	.7365	.4615	.6585
$\epsilon = 0$	23.4153	.7014	.2360	.2011	14.7950	$\epsilon = 0$	.5359	.7295	.4539	.6521
VAE LoRA off	23.0239	.7188	.2458	.2074	17.4872	VAE LoRA off	.5389	.7335	.4572	.6552

(c) Input controls							(d) Repeated stochastic inference		
Control	PSNR	SSIM	LPIPS	DISTS	T-ACC	T-NED	Metric	LoRA	LoRA+LRC
LR input	21.9863	.5799	.6519	.3964	.4235	.6102	PSNR	24.4242 $\pm$.0019	24.8082 $\pm$.0013
Frozen VAE	15.5357	.5033	.5913	.4025	.3427	.5245	SSIM	.710845 $\pm$.000034	.723514 $\pm$.000012
							LPIPS	.241043 $\pm$.000078	.228854 $\pm$.000054
							DISTS	.197032 $\pm$.000057	.195314 $\pm$.000031
							T-ACC	.537603 $\pm$.000143	.540487 $\pm$.001192
							T-NED	.734259 $\pm$.000461	.736198 $\pm$.000637

Table S11: Inference controls and stochastic stability on CTR-TSR-Test $\times 4$. Panels (a)–(c) use fixed checkpoints; panel (d) reports mean $\pm$ standard deviation across repeated inference. VAE LoRA off disables all VAE-encoder LoRA, including the latent projection, while retaining IDM LoRA and LRC.

Comparison	Metric	Change	95% CI
LRC–LoRA	PSNR	+0.3847	[+0.3740, +0.3955]
	SSIM	+0.0127	[+0.0123, +0.0130]
	LPIPS	−0.0121	[−0.0128, −0.0115]
	DISTS	−0.0017	[−0.0021, −0.0013]
	ACC ($p = 0.584$)	+0.0014	[−0.0031, +0.0058]
	NED	+0.0025	[−0.0002, +0.0053]
Pred.–Null	ACC ($p = 0.0035$)	+0.0083	[+0.0028, +0.0138]
	NED	+0.0082	[+0.0051, +0.0113]
GT–Pred.	ACC ($p < 0.001$)	+0.0862	[+0.0797, +0.0925]
	NED	+0.0716	[+0.0675, +0.0757]

Table S12: Filename-paired uncertainty on CTR-TSR-Test $\times 4$. Effect-size intervals use 5,000 paired bootstrap resamples; ACC significance is additionally evaluated with the exact paired McNemar test. Changes are computed from unrounded values.

G.5 Exact Recognition Diagnostics

Tables S13–S16 summarize recognition behavior by confidence, text category, and edit operation. Confidence denotes the mean confidence of the predicted TransOCR tokens. Prior ACC is the exact-match accuracy of the predicted text prior, and D and T denote DiffTSR and TOLA, respectively. The same 8,089 filename-aligned CTR-TSR-Test $\times 4$ samples and TransOCR normalization are used throughout. The confidence relationship is monotonic, but the method improvement is not restricted to high-confidence priors. TOLA improves ACC and NED in all four confidence intervals and all four text categories. The edit decomposition also shows reductions in substitutions, deletions, and insertions, lowering the overall error rate from 42.34 to 32.55 edits per 100 characters. The large gap between correct- and wrong-prior strata identifies prior recognition as the primary remaining source of exact-transcription error.

Confidence	N	Prior	D-A	T-A	D-N	T-N
[0, .25)	246	.0041	.0407	.1382	.1618	.3454
[.25, .50)	1,344	.0074	.0476	.1339	.1976	.3694
[.50, .75)	1,926	.0685	.1506	.2721	.4273	.5771
[.75, 1]	4,573	.7208	.7420	.7923	.9063	.9321

Table S13: Exact recognition by predicted-prior confidence. D/T denote DiffTSR/TOLA; A/N denote ACC/NED.

Category	N	D-A	T-A	D-N	T-N
Chinese	6,946	.4587	.5346	.6366	.7251
English	884	.5204	.5792	.7539	.8098
Mixed	120	.1750	.2917	.5543	.6435
Num./Other	139	.6475	.7266	.8487	.9038

Table S14: Exact recognition by text category. TOLA improves over DiffTSR in every category.

Method	Sub.	Del.	Ins.	Total	Edits/100
DiffTSR	14,657	3,139	673	18,469	42.34
TOLA	12,082	1,529	589	14,200	32.55

Table S15: Character-level edits over 43,622 ground-truth characters under the normalized TransOCR evaluation used for ACC and NED.

Prior	N	Exact GT	Copy	Other
Correct	3,439	3,330	–	109
Wrong	4,650	1,031	376	3,243

Table S16: Behavior conditioned on predicted-prior correctness. Counts sum to all 8,089 CTR-TSR-Test samples.

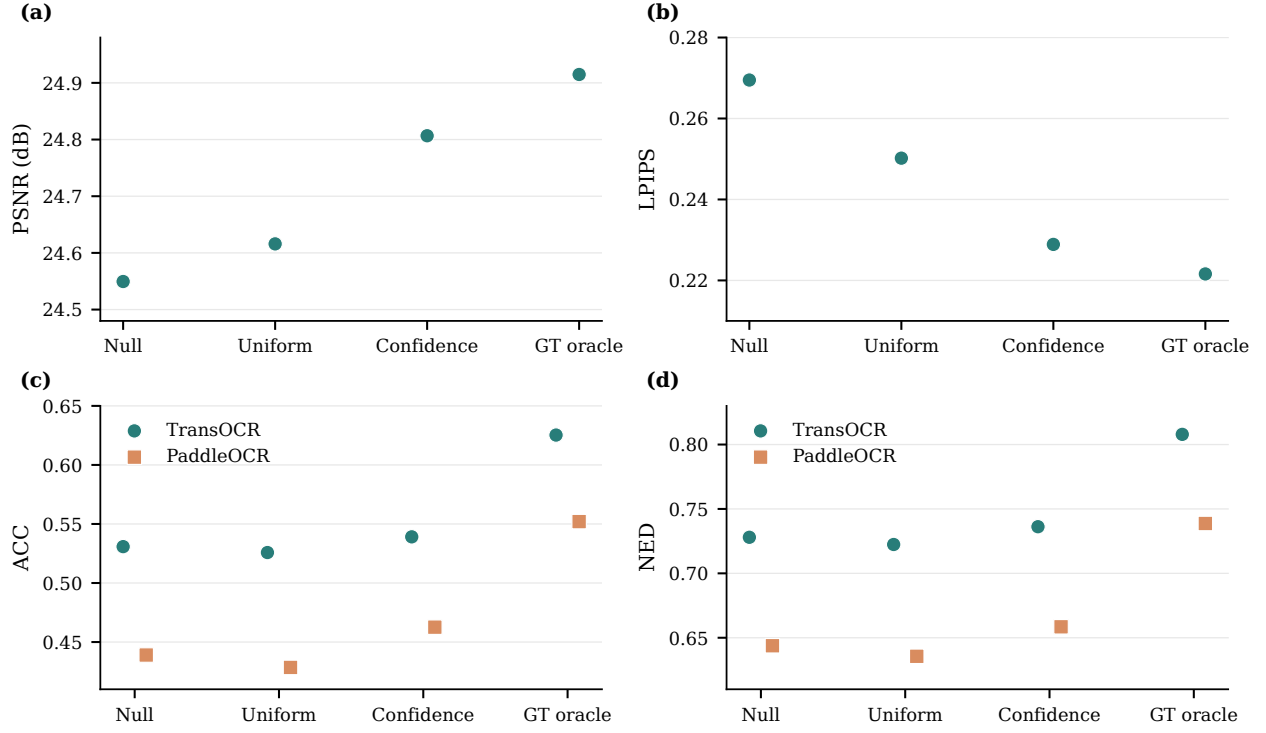


Figure S3: Text-prior controls with null (N), uniform (U), predicted (C), and ground-truth oracle (GT) conditions.

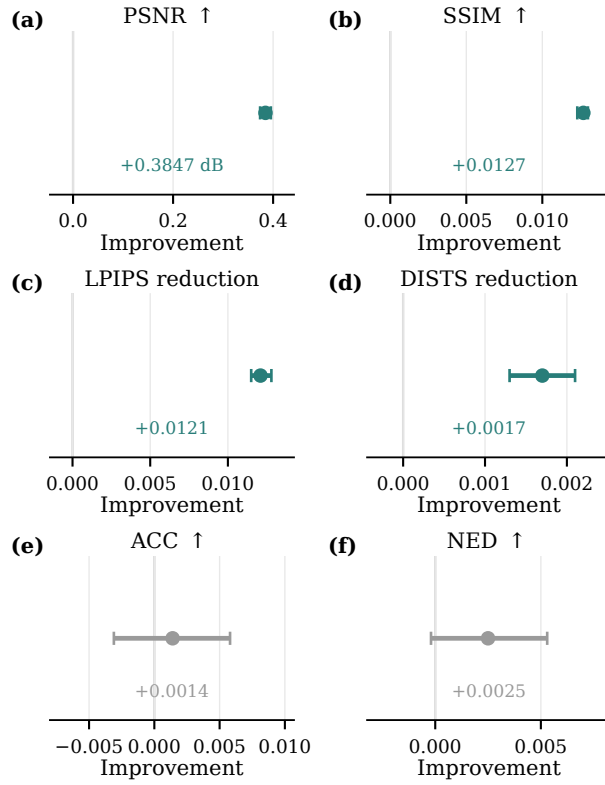


Figure S4: Paired LRC effects over LoRA-only with 95% CIs.