# UNIFIED REINFORCEMENT LEARNING WITH PROGRESSIVE VISUAL REASONING FOR TEXT-AWARE IMAGE RESTORATION

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

## ABSTRACT

Text-Aware Image Restoration is inherently challenging because it is a severely ill-posed problem where character strokes are extremely sensitive to distortion. While recent methods integrate Vision-Language Models for guidance, they often suffer from a VLM drift phenomenon where the guidance policy sacrifices OCR accuracy to satisfy the generative biases of a frozen restoration backbone. In this paper, we propose a novel unified reinforcement learning framework that enables the co-evolution of a VLM-based guidance policy and a Diffusion Transformer restoration model. We formulate the restoration process as an active exploration within the rich pre-trained search space of the generative model to find optimal, text-faithful renderings among multiple plausible solutions. To resolve the non-stationary transition dynamics of this joint environment, our framework employs an Asymmetric Hybrid Reinforcement Learning strategy. Specifically, the VLM is optimized via an on-policy approach using Group Relative Policy Optimization, while the Diffusion Transformer is updated via an off-policy approach using DiffusionNFT. Furthermore, we introduce a Progressive Search mechanism that utilizes intermediate denoised estimates and real-time feedback along the denoising trajectory to iteratively refine the textual search space. Extensive experiments on the SA-Text and Real-Text benchmarks demonstrate that our framework achieves new state-of-the-art performance by effectively reducing hallucinations and significantly enhancing functional legibility.

## 1 INTRODUCTION

Generative models have recently achieved remarkable success in general image restoration tasks (Lin et al., 2024; Yu et al., 2024a; Yue et al., 2023; Wang et al., 2024; Wu et al., 2024a; Ai et al., 2024a; Duan et al., 2025; Fei et al., 2023; Kawar et al., 2022; Zhu et al., 2023; Ye et al., 2024; Chen et al., 2025; Mei et al., 2024). However, they continue to face significant challenges in Text-Aware Image Restoration, where the goal is to recover high-fidelity images containing precise textual information such as street signs, building logos, and official documents (Kim et al., 2025; Min et al., 2025; Liu et al., 2023a; Noguchi et al., 2024; Singh et al., 2024; Ye et al., 2025; Hu et al., 2025; Luo et al., 2025a). Unlike general texture synthesis, where minor structural deviations might be perceptually negligible, textual content is extremely sensitive. Even a slight distortion in a single character's stroke can lead to illegibility or completely alter the semantic meaning (Yang et al., 2023; Tuo et al., 2024). Current diffusion-based approaches, which rely primarily on visual priors, often suffer from text hallucinations, producing visually plausible but semantically nonsensical symbols.

To address these limitations, recent frameworks have introduced explicit textual guidance by integrating Vision-Language Models (VLMs) (Bai et al., 2025; Liu et al., 2023b) and Text Spotting Modules (Zhang et al., 2022). However, conventional vision-based reasoning frameworks (Bai et al., 2025; Liu et al., 2023b; Wu et al., 2024c; Alayrac et al., 2022; Chen et al., 2024; Hurst et al., 2024) predominantly operate on static images. The standard paradigm involves translating visual information into discrete textual tokens for purely text-centric reasoning. This unidirectional abstraction is fundamentally insufficient for generative tasks like Text-Aware Image Restoration. By nature, image restoration is a severely ill-posed problem where a single degraded input can map to multiple plausible solutions. Therefore, solving this task requires more than just extracting a correct text



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

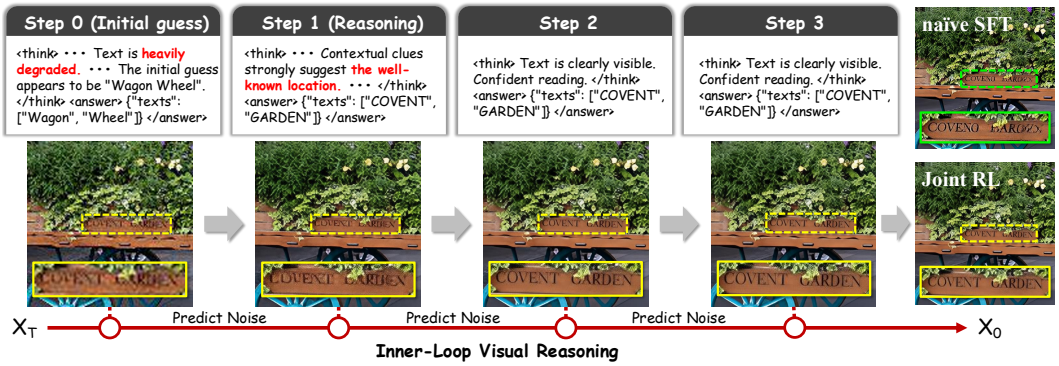



Figure 1: Illustration of the Progressive Search mechanism and a qualitative comparison between the naïve SFT approach and our Joint RL framework. As the denoising trajectory progresses (Step 0 to Step 3), the vision-language model (VLM) iteratively refines its reasoning and textual guidance by interacting with the intermediate visual evidence. While the naïve SFT baseline suffers from VLM drift and struggles to recover accurate character strokes, our co-evolutionary Joint RL framework effectively mitigates hallucinations and restores precise, functionally legible typography.

answer. It necessitates an active exploration within the rich pre-trained search space of a generative model to find the optimal rendering that is faithful to the text among many viable solutions.

This represents a paradigm shift from passive, static observation to dynamic, denoising-aware visual reasoning. Because the problem requires navigating a vast generative space, applying conventional explicit guidance naively leads to a critical VLM drift phenomenon (Zhai et al., 2023; Gao et al., 2023; Ouyang et al., 2022). This typically happens when the restoration backbone remains frozen or is optimized in isolation. The guidance policy (Shao et al., 2024; Schulman et al., 2017) of the VLM tends to sacrifice its intrinsic OCR accuracy to satisfy the generative biases of the frozen backbone. This highlights the urgent need for a customized reinforcement learning framework where both the reasoning policy and the rendering trajectory can jointly evolve.

In this paper, we propose a novel Joint Reinforcement Learning framework where the VLM policy and the diffusion restoration backbone jointly evolve to achieve high-fidelity text reconstruction. Our training pipeline consists of two pivotal stages. First, we conduct Supervised Fine-Tuning on the VLM using Chain-of-Thought data. This stage is crucial not only for grounding the model in visual text understanding but also for enabling the policy to master the required reasoning format. Through Supervised Fine-Tuning, the VLM learns to process OCR hints, perform step-by-step analysis within designated reasoning tags, and produce strictly structured outputs, which is a prerequisite for the subsequent reinforcement learning phase.

Second, we unfreeze the diffusion model parameters and perform joint optimization with the VLM. To overcome the theoretical intractability and non-stationary transition dynamics caused by the continuously shifting generative guidance, we propose an Asymmetric Hybrid Reinforcement Learning strategy. The VLM is updated via an on-policy approach using Group Relative Policy Optimization (Shao et al., 2024), while the diffusion model is updated via an off-policy approach using DiffusionNFT (Zheng et al., 2025). A key innovation in our framework is the Progressive Search mechanism. During the denoising process, the VLM iteratively observes the low-quality input, the intermediate denoised estimate, and real-time feedback along the timesteps. By interacting with the emerging visual evidence, the VLM refines its guidance in a diverse search space, while the diffusion model simultaneously learns to translate these cues into sharp, accurate character strokes. This asymmetric joint evolutionary process effectively resolves the VLM drift problem, as both models collaboratively navigate toward the ground-truth text.

Our experimental results demonstrate that our joint approach significantly reduces hallucinations and outperforms previous models in both functional legibility and perceptual fidelity. Our primary contributions are summarized as follows:

- We present the first **Asymmetric Joint RL framework** for Text-Aware Image Restoration, a hybrid optimization strategy coupling on-policy VLM updates with off-policy diffusion flow matching to ensure stable co-evolution in a non-stationary environment.

- We introduce a **Progressive Search mechanism** integrated into a hierarchical training pipeline, leveraging format-aware Chain-of-Thought SFT and intermediate denoised estimates to iteratively refine the textual search space along the denoising trajectory.
- We achieve new **state-of-the-art performance** across multiple benchmarks, demonstrating superior stability and text reconstruction accuracy compared to existing methods.

## 2 Related works

**Text-aware image restoration.** While early image restoration typically prioritized global visual quality over the fine details of scene text, this gap led to the emergence of Scene Text Image Super-Resolution alongside its standard benchmarks (Karatzas et al., 2015; Risnumawan et al., 2014; Luo et al., 2019; Wang et al., 2020; Shi et al., 2016). Methodologies steadily advanced from foundational CNNs and RNNs to sophisticated GANs and Attention mechanisms (Mou et al., 2020; Cheng et al., 2018; Luo et al., 2019; Liu et al., 2018b;a; Liao et al., 2019; Ma et al., 2022). The subsequent advent of diffusion models (Rombach et al., 2022; Esser et al., 2024; Peebles & Xie, 2023) revolutionized the field by providing powerful generative priors for both general image restoration (Ai et al., 2024b; Chen et al., 2025; Duan et al., 2025; Mei et al., 2024; Lin et al., 2024) and text recognition (Liu et al., 2023a; Noguchi et al., 2024; Singh et al., 2024; Ye et al., 2025; Hu et al., 2025; Luo et al., 2025a). Building upon this foundation, Min et al. (2025) introduced Text-Aware Image Restoration as a comprehensive task aiming for both precise text recovery and high visual quality. To tackle this task, they proposed a multi-task diffusion framework that jointly trains a text-spotting module and leverages its intermediate text predictions as prompts to explicitly guide the subsequent denoising steps. Furthermore, to achieve high-fidelity text restoration, Kim et al. (2025) employ a VLM (Bai et al., 2025; Liu et al., 2023b), utilizing their deep semantic reasoning to actively correct inaccurate text predictions and guide the restoration process.

**VLMs for image restoration.** Empowered by world knowledge captured from massive image-text corpora, VLMs (Bai et al., 2025; Liu et al., 2023b; Hurst et al., 2024) have demonstrated capabilities in fine-grained visual perception and semantic reasoning. This semantic understanding has advanced image restoration from passive, pixel-level mapping toward active, semantic-driven recovery. Alongside advancements in diffusion architectures, recent methods integrate VLMs as interpreters and context correctors within restoration pipelines. For instance, Yu et al. (2024a); Conde et al. (2024); Jiang et al. (2024) extract instructions regarding image content and degradation from VLMs to guide restoration. Furthermore, Zhu et al. (2024) introduce an agentic tool-calling paradigm to orchestrate specialized restoration models based on degradation feedback, while Wang et al. (2025); Luo et al. (2025b); Lin et al. (2025) utilize VLMs to explicitly extract degradation profiles.

**Reinforcement learning (RL)** has become pivotal for aligning generative models with complex human preferences (Black et al., 2023; Wallace et al., 2024; Yang et al., 2024; Liang et al., 2024; Chatterjee et al., 2024; Wu et al., 2024b; Yuan et al., 2024). While standard algorithms like PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024) are widely adopted, their application to continuous diffusion trajectories incurs severe computational bottlenecks. Recent frameworks address this by adapting RL to stochastic flows (Liu et al., 2025; Xue et al., 2025) or employing likelihood-free flow matching (Zheng et al., 2025). Concurrently, PromptLoop (Lee & Ye, 2025) redefines the diffusion Markov Decision Process (Puterman, 2014), treating intermediate latents as states and dynamic textual conditions as actions to steer generation. However, despite these advancements, existing RL-driven diffusion frameworks rely on a stationary diffusion model to bypass the non-Markovian likelihood intractability arising from concurrently optimizing a dynamic guiding policy and the underlying diffusion network. Consequently, the literature still lacks a viable framework capable of jointly optimizing an active reasoning policy alongside the diffusion velocity field, a synergistic co-evolution required to tackle complex conditional generation tasks.

## 3 Method

### 3.1 Overview

Text-Aware Image Restoration requires the precise reconstruction of severely degraded images while maintaining high-frequency textual structures. Conventional text-guided restoration methods (Kim

et al., 2025; Min et al., 2025) typically rely on unidirectional, static text hints. While recent works introduce step-wise prompt refinement, they inherently assume a stationary Markov Decision Process (MDP) by keeping the diffusion model strictly frozen, which fundamentally limits the potential for synergistic adaptation between the visual and language modalities.

To overcome this limitation, we formulate the dynamic interaction between textual reasoning and visual rendering as a *Co-evolving Non-Stationary MDP*. We propose an **Asymmetric Hybrid RL** framework that jointly optimizes a Vision-Language Model (VLM) (Bai et al., 2025) acting as a dynamic navigator and a Diffusion Transformer (DiT) (Duan et al., 2025) serving as the continuous visual renderer. The overall pipeline is systematically structured into two primary phases.

**Phase I: Format-Aware Policy Initialization via SFT.** Before the joint RL, we perform Supervised Fine-Tuning on the VLM to instill a structured, format-aware reasoning grammar (i.e., step-by-step Chain-of-Thought utilizing `<think>` and `<answer>` tags) (Guo et al., 2025; Wei et al., 2022). This phase equips the VLM with an anti-hallucination prior and visual evidence prioritization, acting as a robust cold-start policy for the subsequent RL phase (Yu et al., 2024b; Guo et al., 2025).

**Phase II: Asymmetric Joint RL.** Driven by a shared cooperative reward evaluated on the final restored image, the two models undergo a closed-loop co-evolution. During the rollout phase, we introduce an *Asymmetric Sampling* strategy: the VLM performs stochastic decoding to explore the discrete text condition space, while the DiT employs a deterministic ODE solver (DPM-Solver-2 (Lu et al., 2025)) to isolate credit assignment and preserve high-frequency text strokes. Through *Progressive Search*, the VLM dynamically refines its prompts by continuously observing the estimated clean latent ($\hat{x}_0$) at each denoising timestep. To resolve the non-stationarity caused by this dynamic guidance, we decouple the optimization processes: the VLM is updated via strictly on-policy Group Relative Policy Optimization (GRPO) (Shao et al., 2024) anchored to the frozen SFT policy, whereas the DiT is updated via off-policy DiffusionNFT (Zheng et al., 2025) (forward-process flow matching). This asymmetric architecture successfully prevents trajectory likelihood collapse, ensuring stable convergence within the multi-agent cooperative game.

### 3.2 Phase I: Format-Aware Policy Initialization via SFT

Before engaging in joint reinforcement learning, we perform Supervised Fine-Tuning (SFT) to endow the VLM with a structured visual reasoning grammar and a hallucination-free OCR prior. We adopt a teacher-student distillation approach, generating step-by-step Chain-of-Thought (CoT) traces using a larger teacher model (Qwen2.5-VL-32B) under a deterministic setting to train our 7B-parameter student policy. This initialization is crucial because the subsequent RL phase heavily relies on the VLM's ability to output precisely formatted text conditioning for the DiT.

**Format-Aware Reasoning Grammar.** To structure the VLM's cognitive process, we enforce a strict output grammar comprising a `<think>` block followed by an `<answer>` block with a JSON schema. The reasoning depth is adaptive to the degradation level of the image. For clearly visible text, the model generates a brief confirmation. Conversely, for blurry or truncated text, the model is trained to perform a detailed 4-step reasoning process: identifying the trigger (unclear parts), listing visible evidence (clear strokes), checking plausibility based on context, and drawing a safe conclusion. This "visual evidence first" rule prevents wild guessing and ensures the RL reward formulation operates on a structurally sound foundation.

**Multi-Source Curriculum for Progressive Search.** During the RL phase, the VLM encounters various hint states ranging from empty hints at the initial noise level to rich detections at later steps. To cover this dynamic distribution, our SFT dataset is constructed via a 4-mode curriculum. This includes hint verification with external text spotters, cold-reading without any hints for both clean and severely degraded images, and adversarial hint correction where ground-truth (GT) words are intentionally corrupted with visually similar characters (e.g., swapping 'O' with '0'). The adversarial mode explicitly teaches the policy to trust its visual observations over erroneous textual hints. Furthermore, we utilize high-quality (HQ) and degraded (LQ) image pairs to ensure the student remains robust against the deterministic degradations encountered during RL rollouts.

**Strict Filtering and the KL Anchor.** To guarantee a hallucination-free prior, we apply an aggressive GT-strict filtering mechanism, retaining only the teacher's outputs where the parsed words perfectly match the GT set. The resulting SFT policy not only demonstrates significant improve-

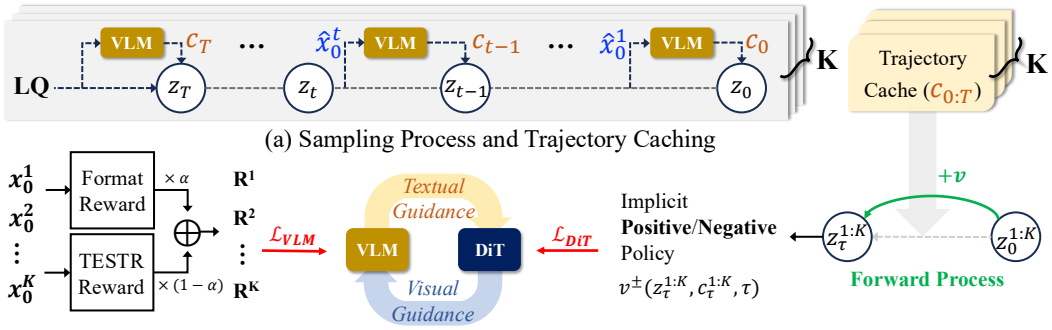



Figure 2: **Overview of the Joint RL framework.** (a) Sampling & Caching: The VLM iteratively observes intermediate latents $\hat{x}_0^{(t)}$ to refine textual guidance $c_t$, while the DiT deterministically renders the image. (b) VLM Update: The VLM is updated strictly on-policy via GRPO using format and OCR (TESTR) rewards. (c) DiT Update: To resolve non-stationarity, the DiT is updated off-policy via DiffusionNFT (forward-process flow matching) using the cached trajectories.

ments in standalone text spotting capabilities but also serves a critical architectural role in Phase II. Specifically, this fine-tuned LoRA adapter (Hu et al., 2022) is retained as a frozen reference policy ($\pi_{\text{sft_ref}}$) during the GRPO update. By anchoring the Kullback-Leibler (KL) divergence penalty to this frozen SFT model, we prevent the VLM from deviating from the established reasoning grammar and drifting into reward hacking during the non-stationary joint evolution.

### 3.3 Problem Formulation: From Stationary to Co-evolving MDP

Following PromptLoop (Lee & Ye, 2025), the step-wise prompt refinement process can be formulated as a Markov Decision Process (MDP) defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$. At each denoising timestep $t$, the state is defined as $s_t = (\hat{x}_0^{(t)}, c_t, q, t)$, where $\hat{x}_0^{(t)}$ is the estimated clean latent and $q$ is the user query. The VLM acts as the policy $\pi_\phi$, sampling an action $a_t = c_{t-1} \sim \pi_\phi(\cdot|s_t)$ which corresponds to the refined prompt. The environment transition $\mathcal{P}(s_{t-1}|s_t, a_t)$ is governed by the DiT velocity field $v_\theta$, and the terminal reward is $R = r(x_0)$ evaluated on the final restored image.

Previous works inherently assumed a *stationary* MDP by keeping the diffusion model strictly frozen ($v_\theta = \text{const}$). However, to achieve true synergistic co-evolution, we must concurrently optimize the DiT ($v_\theta$). This transition introduces two critical theoretical bottlenecks. First, as the DiT updates its parameters, the transition dynamics $\mathcal{P}$ continually shift, creating a Moving Target problem that violates the stationarity assumption of classical RL algorithms.

Second, and more fundamentally, standard diffusion RL (e.g., FlowGRPO (Liu et al., 2025)) optimizes the reverse trajectory likelihood. For a given prompt sequence $c_{0:T}$, the probability of the reverse sampling trajectory $\tau$ is defined as:

$$\log p_\theta(\tau|q) = \log p(z_T) + \sum_{t=1}^{T} \log p_\theta(z_{t-1}|z_t, c_{t-1}) \tag{1}$$

When the condition $c_{t-1}$ is dynamically generated by the VLM at each step based on the intermediate state $\hat{x}_0^{(t)}$, the condition itself becomes a random variable dependent on the current velocity field $v_\theta$. This renders the transition non-Markovian with respect to the initial condition, making the exact calculation of the sequence likelihood intractable and leading to optimization collapse.

### 3.4 Asymmetric Hybrid RL: Fusion of On-Policy and Off-Policy

To break these intractability bottlenecks, we propose an *Asymmetric Hybrid RL* framework that decouples the optimization strategies, fusing an On-Policy approach for the VLM with an Off-Policy approach for the DiT.

**VLM: On-Policy Textual Guidance via GRPO.** To survive the non-stationary environment, the VLM requires strictly *on-policy* updates utilizing only the most recent rollout trajectories. During the rollout, we generate $K = 8$ candidate trajectories using a deterministic ODE solver (DPM-Solver-2 (Lu et al., 2025)) for the DiT. This determinism ensures the variance in the cooperative reward $r(x_0)$ is exclusively attributed to the VLM's stochastic prompt distribution.

We optimize the VLM via Group Relative Policy Optimization (GRPO). For each candidate $i$, we compute a per-group z-score advantage $A_i$:

$$A_i = \frac{r_i - \mu(r_{1:K})}{\sigma(r_{1:K}) + \epsilon}, \quad \text{where} \quad r_i = R_{\text{uni}} \tag{2}$$

$R_{\text{uni}}$ is a shared reward function described in the next stage. To prevent erratic drift, we anchor the Kullback-Leibler (KL) penalty to the frozen SFT policy ($\pi_{\text{sft_ref}}$). The length-normalized log-likelihood for the generated token sequence $y$ is computed as:

$$\bar{\ell}_{\text{curr}}(s_t) = \frac{1}{|y|} \sum_{k=1}^{|y|} \log \pi_\phi(y^{(k)} | \hat{x}_0^{(t)}, h_t, y^{(<k)}) \tag{3}$$

The strictly on-policy VLM objective is then formulated to maximize the expected advantage constrained by the KL trust region:

$$\mathcal{L}_{\text{VLM}}(\phi) = \frac{1}{N} \sum_s \left[ -A_s \cdot \bar{\ell}_{\text{curr}}(s) + \beta_{\text{VLM}} \left( \bar{\ell}_{\text{curr}}(s) - \bar{\ell}_{\text{ref}}(s) \right) \right] \tag{4}$$

**DiT: Off-Policy Image Restoration via DiffusionNFT.** To update the DiT without the intractable reverse trajectory likelihood caused by the dynamic $c_t$, we shift the RL objective to the forward process via DiffusionNFT (Zheng et al., 2025). We construct the forward-process state $z_\tau$ using the final restored latent $z_0$ and artificial noise $\epsilon$:

$$z_\tau = (1 - \tau) z_0 + \tau \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{5}$$

We then formulate the implicit positive and negative policies using a mixing coefficient $\beta_{\text{NFT}}$ and the historical policy $v_{\text{old}}$:

$$v^+ = \beta_{\text{NFT}} v_\theta + (1 - \beta_{\text{NFT}}) v_{\text{old}}, \quad v^- = (1 + \beta_{\text{NFT}}) v_{\text{old}} - \beta_{\text{NFT}} v_\theta \tag{6}$$

The DiT is updated off-policy by minimizing the forward-process flow matching objective. We define the adaptively weighted reconstruction error $\mathcal{L}^\pm$ as:

$$\mathcal{L}^\pm(\theta) = \mathbb{E}_{\text{spatial}} \left[ \frac{\| (z_\tau - \tau v^\pm(z_\tau, c_\tau, \tau)) - z_0 \|_2^2}{w^\pm} \right] \tag{7}$$

The final objective is scaled by the soft positive ratio $\tilde{r} \in [0, 1]$ mapped from the clipped advantage:

$$\mathcal{L}_{\text{DiT}}(\theta) = \left( \frac{\tilde{r}}{\beta_{\text{NFT}}} \mathcal{L}^+ + \frac{1 - \tilde{r}}{\beta_{\text{NFT}}} \mathcal{L}^- \right) \cdot A_{\max} \tag{8}$$

Crucially, $\mathcal{L}_{\text{DiT}}$ is conditionally independent of the reverse sampling sequence, depending only on the marginal tuple $(z_0, c_\tau, r)$. Through *Trajectory-Aware Prompt Replay*, we inject the per-timestep VLM prompts $c_\tau$ cached during the rollout time-aligned with the training noise level $\tau$. This off-policy formulation gracefully bypasses the likelihood intractability while explicitly conditioning the DiT on the VLM's progressive cognitive trajectory, completing the robust co-evolutionary loop. The complete procedure for our proposed co-evolutionary training strategy, including the asymmetric sampling and hybrid optimization steps, is summarized in Algorithm 1 in supplementary material.

**Shared Reward Function.** To align the DiT and VLM toward a unified objective, we propose a Shared Reward $R_{\text{uni}}$ that integrates internal textual perception with external visual verification:

$$R_{\text{uni}} = \alpha \cdot (R_{\text{vlm}} \cdot \mathbb{1}_{\text{fmt}}) + (1 - \alpha) \cdot R_{\text{dit}} \tag{9}$$

where $\alpha = 0.7$. Here, $R_{\text{dit}}$ is an objective score derived from a standalone TESTR model (Zhang et al., 2022) ($0.6F_{\text{det}} + 0.2F_{\text{none}} + 0.2F_{\text{full}}$), while $R_{\text{vlm}}$ represents the word-level recall of the VLM. Crucially, the Format Gate $\mathbb{1}_{\text{fmt}}$ acts as a binary indicator that enforces the structural and qualitative integrity of the CoT reasoning. A violation of the predefined `<think>/<answer>` sequence or a lack of necessary reasoning markers results in $\mathbb{1}_{\text{fmt}} = 0$, effectively nullifying 70% of the total potential reward. This mechanism imposes an implicit penalty that preserves the VLM's logical framework while grounding the entire system's progress in verifiable visual evidence.

Table 1: Quantitative evaluation on the SA-Text test set (Min et al., 2025) across three degradation levels, where degradation becomes progressively more severe towards Level 3. Best and second-best performances are indicated in **bold** and underline, respectively.

| Deg. Level | Model | ABCNet v2 (Liu et al., 2021) | | | | | | TESTR (Zhang et al., 2022) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Detection | | | End-to-End | | | Detection | | | End-to-End | | |
| | | Precision(↑) | Recall(↑) | F1-Score(↑) | None(↑) | Full(↑) | Avg(↑) | Precision(↑) | Recall(↑) | F1-Score(↑) | None(↑) | Full(↑) | Avg(↑) |
| - | HQ (GT) | 92.16 | 86.85 | 89.43 | 71.79 | 82.05 | 76.92 | 92.59 | 87.81 | 90.13 | 75.90 | 84.18 | 80.04 |
| Level1 | LQ (Lv1) | 89.79 | 29.51 | 44.42 | 24.29 | 34.25 | 29.27 | 84.01 | 30.24 | 44.47 | 25.93 | 34.73 | 30.33 |
| | Real-ESRGAN (Wang et al., 2021) | 83.79 | 43.34 | 57.13 | 21.45 | 30.12 | 25.79 | 85.19 | 41.98 | 56.24 | 22.90 | 31.52 | 27.21 |
| | SwinIR (Liang et al., 2021) | 84.95 | 40.93 | 55.25 | 22.70 | 31.26 | 26.98 | 87.93 | 39.62 | 54.63 | 25.50 | 33.75 | 29.63 |
| | ResShift (Yue et al., 2023) | 81.93 | 40.07 | 53.82 | 20.40 | 28.74 | 24.57 | **88.47** | 35.81 | 50.98 | 22.39 | 30.14 | 26.27 |
| | StableSR (Wang et al., 2024) | 77.90 | 55.44 | 64.78 | 21.93 | 29.29 | 25.61 | 84.44 | 50.68 | 63.34 | 24.02 | 31.84 | 27.93 |
| | DiffBIR (Lin et al., 2024) | 76.29 | 56.44 | 64.88 | 23.14 | 32.73 | 27.94 | 84.00 | 52.13 | 64.34 | 25.51 | 35.47 | 30.49 |
| | SeeSR (Wu et al., 2024a) | 70.00 | **61.88** | 65.69 | 20.16 | 28.63 | 24.40 | 78.85 | 55.76 | 65.32 | 23.31 | 32.82 | 28.07 |
| | SUPIR (Yu et al., 2024a) | 43.64 | 49.46 | 46.37 | 14.58 | 19.34 | 16.96 | 53.02 | 46.19 | 49.37 | 17.44 | 22.29 | 19.87 |
| | FaithDiff (Chen et al., 2025) | 69.16 | 61.51 | 65.12 | 20.44 | 27.78 | 24.11 | 78.80 | **57.12** | 66.23 | 22.50 | 31.59 | 27.05 |
| | TeReDiff (Min et al., 2025) | 85.29 | 58.34 | **69.29** | 26.59 | 35.69 | 31.14 | 87.50 | 54.90 | 67.47 | 28.19 | 36.99 | 32.59 |
| | DiT4SR (Duan et al., 2025) | 73.68 | 58.11 | 64.98 | 26.56 | 34.31 | 30.44 | 81.47 | 54.81 | 65.53 | 28.94 | 36.48 | 32.71 |
| | UniT (Kim et al., 2025) | 85.89 | 55.49 | 67.42 | 30.96 | 39.93 | 35.45 | 87.47 | 51.59 | 64.90 | 31.99 | 41.52 | 36.76 |
| | Ours | **85.99** | 58.03 | **69.29** | **35.58** | **43.46** | **39.52** | 87.88 | 56.18 | **68.54** | **37.82** | **45.47** | **41.65** |
| Level2 | LQ (Lv2) | 87.67 | 22.89 | 36.30 | 20.49 | 27.82 | 24.16 | 78.45 | 23.93 | 36.68 | 20.49 | 27.37 | 23.93 |
| | Real-ESRGAN (Wang et al., 2021) | 81.42 | 41.12 | 54.64 | 18.31 | 24.88 | 21.60 | 84.92 | 38.80 | 53.27 | 19.29 | 27.50 | 23.40 |
| | SwinIR (Liang et al., 2021) | 80.14 | 37.31 | 50.91 | 17.82 | 24.93 | 21.38 | 85.43 | 34.81 | 49.47 | 19.07 | 26.99 | 23.03 |
| | ResShift (Yue et al., 2023) | 81.11 | 35.22 | 49.12 | 17.89 | 25.54 | 21.72 | 85.18 | 32.05 | 46.57 | 17.26 | 26.09 | 21.68 |
| | StableSR (Wang et al., 2024) | 75.49 | 51.95 | 61.55 | 19.55 | 26.69 | 23.12 | 79.03 | 48.87 | 60.39 | 20.06 | 27.68 | 23.87 |
| | DiffBIR (Lin et al., 2024) | 72.96 | 53.94 | 62.03 | 19.60 | 27.52 | 23.56 | 79.69 | 50.50 | 61.82 | 21.64 | 30.36 | 26.00 |
| | SeeSR (Wu et al., 2024a) | 68.93 | **60.65** | 64.53 | 19.48 | 26.72 | 23.10 | 77.50 | **54.49** | 63.99 | 21.83 | 29.17 | 25.50 |
| | SUPIR (Yu et al., 2024a) | 42.01 | 45.65 | 43.75 | 13.21 | 17.42 | 15.32 | 53.54 | 43.25 | 47.84 | 15.50 | 19.96 | 17.73 |
| | FaithDiff (Chen et al., 2025) | 66.62 | 59.34 | 62.77 | 18.94 | 25.99 | 22.47 | 76.16 | 54.17 | 63.31 | 20.98 | 28.56 | 24.77 |
| | TeReDiff (Min et al., 2025) | 83.02 | 56.30 | **67.10** | 24.42 | 33.23 | 28.83 | 86.95 | 52.86 | **65.75** | 26.39 | 35.13 | 30.76 |
| | DiT4SR (Duan et al., 2025) | 69.22 | 54.03 | 60.69 | 23.27 | 29.99 | 26.63 | 76.56 | 51.09 | 61.28 | 25.39 | 32.63 | 29.01 |
| | UniT (Kim et al., 2025) | 83.73 | 50.86 | 63.28 | 28.43 | 36.83 | 32.63 | 87.53 | 48.69 | 62.57 | 30.12 | 37.98 | 34.05 |
| | Ours | **85.08** | 54.06 | 66.11 | **30.45** | **37.33** | **33.89** | **87.88** | 52.29 | 65.57 | **32.07** | **39.12** | **35.60** |
| Level3 | LQ (Lv3) | 85.38 | 13.24 | 22.92 | 12.17 | 16.95 | 14.56 | 76.01 | 15.37 | 25.57 | 12.52 | 17.72 | 15.12 |
| | Real-ESRGAN (Wang et al., 2021) | 72.48 | 28.65 | 41.07 | 11.89 | 16.37 | 14.13 | 76.51 | 27.02 | 39.93 | 12.13 | 18.22 | 15.18 |
| | SwinIR (Liang et al., 2021) | 74.41 | 25.57 | 38.06 | 11.27 | 16.33 | 13.80 | 78.38 | 24.16 | 36.94 | 11.85 | 17.12 | 14.49 |
| | ResShift (Yue et al., 2023) | 75.00 | 22.57 | 34.70 | 10.80 | 15.19 | 13.00 | 81.10 | 20.04 | 32.13 | 9.89 | 15.63 | 12.76 |
| | StableSR (Wang et al., 2024) | 67.63 | 38.08 | 48.72 | 13.34 | 18.91 | 16.13 | 72.21 | 35.22 | 47.35 | 13.65 | 19.68 | 16.67 |
| | DiffBIR (Lin et al., 2024) | 59.30 | 42.20 | 49.31 | 13.88 | 19.39 | 16.64 | 72.27 | 38.98 | 50.65 | 15.61 | 22.67 | 19.14 |
| | SeeSR (Wu et al., 2024a) | 55.06 | 46.83 | 50.61 | 13.38 | 18.47 | 15.93 | 64.95 | 43.93 | 52.41 | 14.93 | 20.88 | 17.91 |
| | SUPIR (Yu et al., 2024a) | 31.05 | 34.72 | 32.78 | 9.07 | 11.77 | 10.42 | 40.78 | 32.77 | 36.34 | 11.21 | 14.02 | 12.62 |
| | FaithDiff (Chen et al., 2025) | 56.04 | **47.91** | 51.66 | 13.69 | 19.01 | 16.35 | 69.44 | **45.01** | 54.62 | 15.40 | 21.18 | 18.29 |
| | TeReDiff (Min et al., 2025) | 81.76 | 44.11 | **57.30** | 19.61 | 27.50 | 23.56 | 84.50 | 42.02 | 56.13 | 19.92 | 28.34 | 24.13 |
| | DiT4SR (Duan et al., 2025) | 60.67 | 43.29 | 50.53 | 16.61 | 22.28 | 19.45 | 70.57 | 40.44 | 51.41 | 17.98 | 24.55 | 21.27 |
| | UniT (Kim et al., 2025) | 83.14 | 38.21 | 52.36 | 21.12 | 28.63 | 24.88 | **86.43** | 36.94 | 51.76 | 22.86 | **29.53** | **26.20** |
| | Ours | **83.68** | 42.09 | 56.00 | **22.09** | **29.03** | **25.56** | 83.84 | 42.59 | **56.48** | **23.22** | 28.45 | 25.84 |

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Datasets and evaluation protocol.** Following TeReDiff (Min et al., 2025), we utilize its benchmark datasets for both training (SFT and GRPO) and evaluation. We evaluate on SA-Text test with three levels of synthetic degradation and Real-Text, a collection of real-world HR-LR pairs from RealSR (Cai et al., 2019) and DRealSR (Pengxu Wei, 2020).

**Evaluation metrics.** To quantify text restoration performance, we use off-the-shelf text-spotting models (Liu et al., 2021; Zhang et al., 2022) and report detection Precision, Recall, and F1-score. We also evaluate the end-to-end (E2E) recognition F1-score under two settings: E2E-None, which tests lexicon-free character recovery, and E2E-Full, which assesses recognition using a predefined full lexicon. To assess perceptual image quality, we additionally report LPIPS (full-reference) and NIQE, MUSIQ, CLIP-IQA, MANIQA, and BRISQUE (no-reference).

**Baselines.** To evaluate the effectiveness of our method, we compare it against two primary categories of baselines: conventional image restoration pipelines (Duan et al., 2025; Chen et al., 2025; Yu et al., 2024a; Wu et al., 2024a; Lin et al., 2024; Wang et al., 2024; Yue et al., 2023; Liang et al., 2021; Wang et al., 2021) and recent text-aware image restoration methods (Kim et al., 2025; Min et al., 2025). For these baselines, we directly report the quantitative results provided in (Kim et al., 2025).

**Implementation details.** A Chain-of-Thought dataset of approximately 161k instances was generated using Qwen2.5-VL-32B-Instruct (Bai et al., 2025), followed by Supervised Fine-Tuning (SFT) of Qwen2.5-VL-7B-Instruct (Bai et al., 2025) with LoRA (Hu et al., 2022). For the Asymmetric Joint RL phase, both the VLM and DiT (Kim et al., 2025) are equipped with LoRA adapters and jointly optimized using a shared reward computed by TESTR (Zhang et al., 2022): $0.6 \cdot \text{det_F1} + 0.2 \cdot \text{e2e_none} + 0.2 \cdot \text{e2e_full}$. The DiT is optimized via DiffusionNFT, while the VLM is updated via GRPO with a KL penalty to the frozen SFT adapter. Detailed training configurations and computational costs are provided in the supplementary material.

Table 2: Quantitative evaluation on the Real-Text. Best and second-best performances are indicated in **bold** and underline, respectively.

| Model | ABCNet v2 (Liu et al., 2021) | | | | | | TESTR (Zhang et al., 2022) | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Detection | | | End-to-End | | | Detection | | | End-to-End | | |
| | Precision(↑) | Recall(↑) | F1-Score(↑) | None(↑) | Full(↑) | Avg(↑) | Precision(↑) | Recall(↑) | F1-Score(↑) | None(↑) | Full(↑) | Avg(↑) |
| HQ (GT) | 90.03 | 85.52 | 87.72 | 72.06 | 79.48 | 75.77 | 90.29 | 85.77 | 87.97 | 74.50 | 81.72 | 78.11 |
| LQ | 89.10 | 44.97 | 59.77 | 42.64 | 50.21 | 46.43 | 85.33 | 51.61 | 64.32 | 47.08 | 55.11 | 51.10 |
| Real-ESRGAN (Wang et al., 2021) | 79.15 | 52.70 | 63.27 | 35.30 | 39.88 | 37.59 | 82.67 | 53.94 | 65.29 | 38.16 | 42.36 | 40.26 |
| SwinIR (Liang et al., 2021) | 80.29 | 47.45 | 59.64 | 38.39 | 42.63 | 40.51 | 82.92 | 47.89 | 60.72 | 39.97 | 44.56 | 42.27 |
| ResShift (Yue et al., 2023) | 81.17 | 33.76 | 47.69 | 30.95 | 34.87 | 32.91 | 82.23 | 39.91 | 53.74 | 35.31 | 39.99 | 37.65 |
| StableSR (Wang et al., 2024) | 79.79 | 59.89 | 68.42 | 41.23 | 47.64 | 44.44 | 82.19 | 60.39 | 69.62 | 42.53 | 49.39 | 45.96 |
| DiffBIR (Lin et al., 2024) | 66.04 | 59.69 | 62.71 | 33.75 | 40.05 | 36.90 | 76.33 | 61.87 | 68.35 | 39.27 | 46.11 | 42.69 |
| SeeSR (Wu et al., 2024a) | 68.12 | 63.46 | 65.71 | 37.11 | 43.43 | 40.27 | 74.29 | 62.47 | 67.87 | 40.34 | 46.54 | 43.44 |
| SUPIR (Yu et al., 2024a) | 44.00 | 40.56 | 42.21 | 22.29 | 25.03 | 23.66 | 53.08 | 44.47 | 48.39 | 27.25 | 30.59 | 28.92 |
| FaithDiff (Chen et al., 2025) | 71.21 | 64.50 | 67.69 | 38.81 | 44.28 | 41.55 | 76.90 | 65.20 | 70.57 | 41.64 | 47.97 | 44.81 |
| TeReDiff (Min et al., 2025) | 83.95 | 67.58 | **74.88** | 48.39 | 55.01 | 51.70 | 84.30 | 67.37 | 74.89 | 49.39 | 56.45 | 52.92 |
| DiT4SR (Duan et al., 2025) | 74.94 | **68.77** | 71.72 | 46.17 | 51.67 | 48.92 | 79.53 | **69.25** | 74.04 | 48.93 | 54.58 | 51.76 |
| UniT (Kim et al., 2025) | 85.06 | 60.42 | 70.65 | 49.81 | 56.11 | 52.96 | **86.94** | 62.74 | 72.88 | 53.80 | 59.74 | 56.77 |
| Ours | **87.90** | 63.79 | 73.93 | **53.20** | **58.07** | **55.64** | 86.12 | 67.19 | **75.48** | **55.08** | **61.02** | **58.05** |

Table 3: **Perceptual quality comparison** averaged over SA-Text Levels 1–3 and Real-Text. Bold indicates the best result among the matched four-step methods.

| Method | LPIPS ↓ | NIQE ↓ | MUSIQ ↑ | CLIP-IQA ↑ | MANIQA ↑ | BRISQUE ↓ |
|---|---|---|---|---|---|---|
| UniT (Kim et al., 2025) | 0.355 | 10.81 | 62.70 | 0.562 | 0.513 | 64.27 |
| TeReDiff (Min et al., 2025) | **0.280** | 10.26 | 64.97 | 0.588 | **0.565** | 38.53 |
| Ours | 0.320 | **7.59** | **67.17** | **0.615** | 0.522 | **34.07** |

### 4.2 Comparison with state-of-the-art

**Quantitative comparison.** We evaluate our Joint RL framework against state-of-the-art text-aware and general image restoration methods on both synthetic and real-world datasets. Table 1 reports results on SA-Text across three degradation levels. Our method achieves strong improvements in End-to-End (E2E) recognition, directly reflecting functional legibility and suppression of text hallucinations. Under Level 1 degradation, our framework achieves a TESTR E2E average score of 41.65%, significantly outperforming UniT (36.76%) and TeReDiff (32.59%). On the challenging Real-Text benchmark (Table 2), our model establishes a new state-of-the-art with a score of 58.05%, surpassing UniT (56.77%) and TeReDiff (52.92%). These gains demonstrate that our co-evolutionary optimization effectively aligns VLM guidance with accurate structural restoration while maintaining OCR accuracy under complex real-world degradations. Under extreme Level 3 degradation, where visual evidence is severely diminished, our visual-evidence-first strategy discourages unsupported textual predictions, potentially reducing detection recall and lexicon-assisted recognition scores.

**Perceptual quality.** We evaluate perceptual quality using full- and no-reference metrics. As shown in Table 3, our method achieves favorable no-reference scores, particularly in NIQE, MUSIQ, CLIP-IQA, and BRISQUE. While TeReDiff obtains better LPIPS and MANIQA scores, our method maintains competitive perceptual quality while significantly improving text legibility.

**Qualitative comparison.** Figure 3 demonstrates the qualitative superiority of our Joint RL framework over recent text-aware restoration methods. Although UniT (Kim et al., 2025) and TeReDiff (Min et al., 2025) outperform general restoration baselines, they still suffer from severe text hallucinations and structural distortions under heavy degradation, as shown in the 'LOOSE TEA' and 'FIAT' examples. These decoupled approaches often generate visually plausible yet semantically incorrect symbols because the VLM guidance drifts from the restoration backbone's rendering capability. In contrast, our framework consistently restores sharp, structurally accurate, and semantically faithful typography. Through the co-evolution of the VLM-based guidance policy and the DiT backbone, our model effectively suppresses hallucinations while preserving fine-grained character strokes and functional legibility. The accurate reconstruction of the 'FIAT' logo and the detailed 'LAMBORGHINI' example in Figure 3 highlight the effectiveness of our synergistic optimization in bridging semantic reasoning and high-fidelity visual rendering.

### 4.3 Ablation studies

To validate the effectiveness of each core component within our proposed framework, we conduct an ablation study. Table 4 presents the evaluation of various configurations using the E2E-None and E2E-Full metrics evaluated by ABCNet v2 (Liu et al., 2021) and TESTR (Zhang et al., 2022).

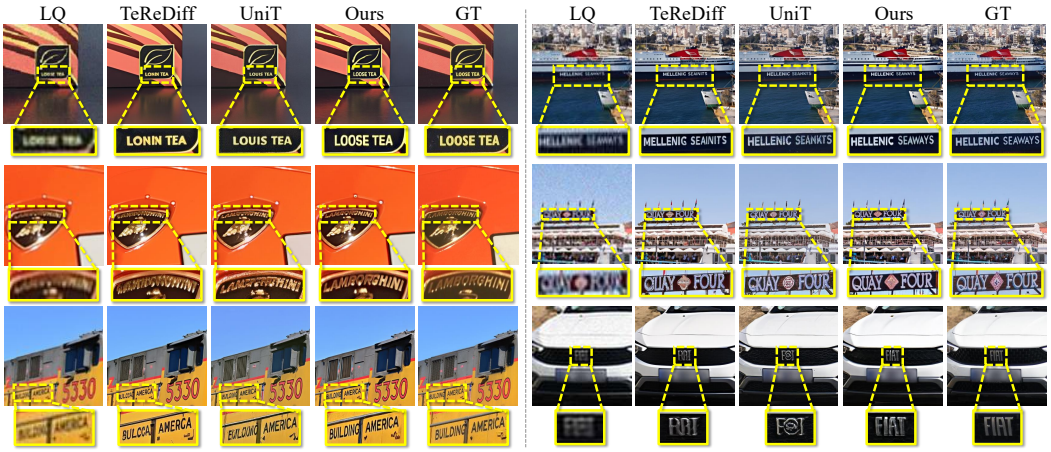



Figure 3: **Qualitative results on the SA-Text and Real-Text.** Visual comparison of our method against low-quality (LQ) inputs, baseline models (TeReDiff (Min et al., 2025), UniT (Kim et al., 2025)), and the ground truth (GT). The text regions are highlighted and zoomed in with yellow boxes for detailed comparison.

Table 4: Ablation study investigating the effectiveness of SFT, RL, and joint co-evolution.

| Setting | SFT | VLM RL | DiT RL | Co-evolution (Joint) | Dynamic State ($\hat{x}_0^{(t)}$) | ABCNet v2 (Liu et al., 2021) | | | TESTR (Zhang et al., 2022) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | None (↑) | Full (↑) | Avg (↑) | None (↑) | Full (↑) | Avg (↑) |
| A | | | | | | 30.96 | 39.93 | 35.45 | 31.99 | 41.52 | 36.76 |
| B | ✓ | | | | | 32.09 | 40.87 | 36.48 | 34.48 | 42.91 | 38.70 |
| C | ✓ | ✓ | | - | ✓ | 31.53 | 41.23 | 36.38 | 33.14 | 42.44 | 37.79 |
| D | ✓ | | ✓ | - | ✓ | <u>35.12</u> | **43.54** | <u>39.33</u> | 37.04 | 45.12 | 41.08 |
| E | ✓ | ✓ | ✓ | ✓ | | 34.93 | 42.11 | 38.52 | <u>37.36</u> | <u>45.41</u> | <u>41.39</u> |
| F (Ours) | ✓ | ✓ | ✓ | ✓ | ✓ | **35.58** | <u>43.46</u> | **39.52** | **37.82** | **45.47** | **41.65** |

**Effectiveness of SFT.** Comparing the baseline (Setting A) with Setting B, we observe that introducing Supervised Fine-Tuning yields a clear performance boost. This confirms that instilling a format-aware reasoning grammar provides a robust cold-start policy, preventing early hallucination issues before the reinforcement learning phase.

**Analysis of VLM Drift and Independent RL.** Setting C applies reinforcement learning strictly to the VLM while keeping the DiT frozen. Interestingly, this unidirectional optimization leads to a performance drop in several metrics compared to Setting B. This empirically corroborates our hypothesis regarding the VLM drift phenomenon. When forced to maximize reward against a stationary backbone, the VLM sacrifices true OCR accuracy to satisfy the generative biases of the frozen model. Conversely, Setting D trains only the DiT using DiffusionNFT. While this yields significant gains over the baseline, it fundamentally lacks the active reasoning capabilities required for complex textual degradations.

**Impact of Joint Co-evolution and Progressive Search.** Setting E introduces the joint co-evolution of both models but omits the dynamic state observation. While it performs strongly, it is ultimately surpassed by Setting F (Ours), which integrates the full Progressive Search mechanism. By allowing the VLM to iteratively observe the intermediate denoised estimate ($\hat{x}_0^{(t)}$), the framework dynamically narrows down the textual search space along the denoising trajectory. This continuous and synergistic interaction culminates in the highest overall performance, demonstrating the absolute necessity of our complete Asymmetric Joint RL architecture.

## 5 Conclusion

We present an **Asymmetric Joint RL** framework that jointly optimizes a Vision-Language Model and a Diffusion Transformer for text-aware image restoration. To handle the intractable likelihoods and non-stationarity of dynamic prompt refinement, the VLM is updated on-policy via GRPO while the DiT is optimized off-policy via DiffusionNFT, conditioned on the per-timestep prompts cached during rollout. Extensive experiments show that this co-evolutionary design reduces text hallucinations and consistently improves end-to-end text recognition on SA-Text and Real-Text.

## REFERENCES


Yuang Ai, Huaibo Huang, Xiaoqiang Zhou, Jiexiang Wang, and Ran He. Multimodal prompt perceiver: Empower adaptiveness generalizability and fidelity for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25432–25444, 2024a.

Yuang Ai, Xiaoqiang Zhou, Huaibo Huang, Xiaotian Han, Zhengyu Chen, Quanzeng You, and Hongxia Yang. Dreamclear: High-capacity real-world image restoration with privacy-safe dataset curation. *Advances in Neural Information Processing Systems*, 37:55443–55469, 2024b.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

Agneet Chatterjee, Gabriela Ben Melech Stan, Estelle Aflalo, Sayak Paul, Dhruba Ghosh, Tejas Gokhale, Ludwig Schmidt, Hannaneh Hajishirzi, Vasudev Lal, Chitta Baral, et al. Getting it right: Improving spatial consistency in text-to-image models. In *European Conference on Computer Vision*, pp. 204–222. Springer, 2024.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Zhanzhan Cheng, Yangliu Xu, Fan Bai, Yi Niu, Shiliang Pu, and Shuigeng Zhou. Aon: Towards arbitrarily-oriented text recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5571–5579, 2018.

Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *European Conference on Computer Vision*, pp. 1–21. Springer, 2024.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9935–9946, 2023.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866. PMLR, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

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. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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. *arXiv preprint arXiv:2506.04641*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Yitong Jiang, Zhaoyang Zhang, Tianfan Xue, and Jinwei Gu. Autodir: Automatic all-in-one image restoration with latent diffusion. In *European Conference on Computer Vision*, pp. 340–359. Springer, 2024.

Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pp. 1156–1160. IEEE, 2015.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35:23593–23606, 2022.

Jin Hyeon Kim, Paul Hyunbin Cho, Claire Kim, Jaewon Min, Jaeeun Lee, Jihye Park, Yeji Choi, and Seungryong Kim. Unified diffusion transformer for high-fidelity text-aware image restoration. *arXiv preprint arXiv:2512.08922*, 2025.

Suhyeon Lee and Jong Chul Ye. Plug-and-play prompt refinement via latent feedback for diffusion model alignment. *arXiv preprint arXiv:2510.00430*, 2025.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2(5):7, 2024.

Minghui Liao, Jian Zhang, Zhaoyi Wan, Fengming Xie, Jiajun Liang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Scene text recognition from two-dimensional perspective. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 8714–8721, 2019.

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.

Yunlong Lin, Zixu Lin, Haoyu Chen, Panwang Pan, Chenxin Li, Sixiang Chen, Kairun Wen, Yeying Jin, Wenbo Li, and Xinghao Ding. Jarvisir: Elevating autonomous driving perception with intelligent image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22369–22380, 2025.

Baolin Liu, Zongyuan Yang, Pengfei Wang, Junjie Zhou, Ziqi Liu, Ziyi Song, Yan Liu, and Yongping Xiong. Textdiff: Mask-guided residual diffusion models for scene text image super-resolution. *arXiv preprint arXiv:2308.06743*, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *arXiv preprint arXiv:2505.05470*, 2025.

Wei Liu, Chaofeng Chen, and Kwan-Yee Wong. Char-net: A character-aware neural network for distorted scene text recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018a.

Yang Liu, Zhaowen Wang, Hailin Jin, and Ian Wassell. Synthetically supervised feature learning for scene text recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 435–451, 2018b.

Yuliang Liu, Chunhua Shen, Lianwen Jin, Tong He, Peng Chen, Chongyu Liu, and Hao Chen. Abcnet v2: Adaptive bezier-curve network for real-time end-to-end text spotting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8048–8064, 2021.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025.

Canjie Luo, Lianwen Jin, and Zenghui Sun. Moran: A multi-object rectified attention network for scene text recognition. *Pattern Recognition*, 90:109–118, 2019.

Minxing Luo, Linlong Fan, Wang Qiushi, Ge Wu, Yiyan Luo, Yuhang Yu, Jinwei Chen, Yaxing Wang, Qingnan Fan, and Jian Yang. Restore text first, enhance image later: Two-stage scene text image super-resolution with glyph structure guidance. *arXiv preprint arXiv:2510.21590*, 2025a.

Wenyang Luo, Haina Qin, Zewen Chen, Libin Wang, Dandan Zheng, Yuming Li, Yufan Liu, Bing Li, and Weiming Hu. Visual-instructed degradation diffusion for all-in-one image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12764–12777, 2025b.

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*, pp. 5911–5920, 2022.

Kangfu Mei, Mauricio Delbracio, Hossein Talebi, Zhengzhong Tu, Vishal M Patel, and Peyman Milanfar. Codi: Conditional diffusion distillation for higher-fidelity and faster image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9048–9058, 2024.

Jaewon Min, Jin Hyeon Kim, Paul Hyunbin Cho, Jaeeun Lee, Jihye Park, Minkyu Park, Sangpil Kim, Hyunhee Park, and Seungryong Kim. Text-aware image restoration with diffusion models. *arXiv preprint arXiv:2506.09993*, 2025.

Yongqiang Mou, Lei Tan, Hui Yang, Jingying Chen, Leyuan Liu, Rui Yan, and Yaohong Huang. Plugnet: Degradation aware scene text recognition supervised by a pluggable super-resolution unit. In *European conference on computer vision*, pp. 158–174. Springer, 2020.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Hannan Lu ZongYuan Zhan Qixiang Ye Wangmeng Zuo Liang Lin Pengxu Wei, Ziwei Xie. Component divide-and-conquer for real-world image super-resolution. In *Proceedings of the European Conference on Computer Vision*, 2020.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

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.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016.

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*, pp. 303–320. Springer, 2024.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. In *International Conference on Learning Representations*, volume 2024, pp. 56783–56799, 2024.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Chao Wang, Hehe Fan, Huichen Yang, Sarvnaz Karimi, Lina Yao, and Yi Yang. Adapting text-to-image generation with feature difference instruction for generic image restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23539–23550, 2025.

Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. *International Journal of Computer Vision*, 132(12):5929–5949, 2024.

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*, pp. 650–666. Springer, 2020.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1905–1914. IEEE, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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, 2024a.

Xindi Wu, Dingli Yu, Yangsibo Huang, Olga Russakovsky, and Sanjeev Arora. Conceptmix: A compositional image generation benchmark with controllable difficulty. *Advances in Neural Information Processing Systems*, 37:86004–86047, 2024b.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024c. URL https://arxiv.org/abs/2412.10302.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using human feedback to fine-tune diffusion models without any reward model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8941–8951, 2024.

Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36:44050–44066, 2023.

Keren Ye, Ignacio Garcia Dorado, Michalis Raptis, Mauricio Delbracio, Irene Zhu, Peyman Milanfar, and Hossein Talebi. Textsr: diffusion super-resolution with multilingual ocr guidance. *arXiv preprint arXiv:2505.23119*, 2025.

Tian Ye, Sixiang Chen, Wenhao Chai, Zhaohu Xing, Jing Qin, Ge Lin, and Lei Zhu. Learning diffusion texture priors for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2524–2534, 2024.

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, 2024a.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024b.

Huizhuo Yuan, Zixiang Chen, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning of diffusion models for text-to-image generation. *Advances in Neural Information Processing Systems*, 37: 73366–73398, 2024.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. *Advances in neural information processing systems*, 36: 13294–13307, 2023.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

Xiang Zhang, Yongwen Su, Subarna Tripathi, and Zhuowen Tu. Text spotting transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9519–9528, 2022.

Kaiwen Zheng, Huayu Chen, Haotian Ye, Haoxiang Wang, Qinsheng Zhang, Kai Jiang, Hang Su, Stefano Ermon, Jun Zhu, and Ming-Yu Liu. Diffusionnft: Online diffusion reinforcement with forward process. *arXiv preprint arXiv:2509.16117*, 2025.

Kaiwen Zhu, Jinjin Gu, Zhiyuan You, Yu Qiao, and Chao Dong. An intelligent agentic system for complex image restoration problems. *arXiv preprint arXiv:2410.17809*, 2024.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1219–1229, 2023.

# A UNIFIED REINFORCEMENT LEARNING ALGORITHM

**Algorithm 1** Asymmetric Joint RL for Text-Aware Image Restoration

**Require:** VLM $\pi_\phi$ (`default`, `sft_ref`), DiT $v_\theta$, Group size $K = 8$, Denoising steps $T = 4$.

**while** training **do**
  **% 1. Asymmetric Rollout** (repeated for $K$ candidates)
  Obtain $x^{\text{LQ}}$ and initialize $z_T \sim \mathcal{N}(0, I)$. Cache $\mathcal{C} \leftarrow \emptyset$.
  **for** $t = T$ **to** $1$ **do**
    $\hat{x}_0^{(t)} \leftarrow \text{EstimateClean}(z_t, t)$
    $c_{t-1} \sim \pi_\phi(\cdot|\hat{x}_0^{(t)}, h_t)$ ▷ *Stochastic VLM Search*
    $\mathcal{C}[t-1] \leftarrow c_{t-1}$ ▷ *Trajectory Caching*
    $z_{t-1} \leftarrow \text{DPM-Solver-2}(z_t, z^{\text{LQ}}, c_{t-1}, t)$ ▷ *Deterministic DiT*
  **end for**
  Decode $x_0 \leftarrow \text{VAE.Decode}(z_0)$.
  **% 2. Cooperative Reward & Advantage**
  Compute rewards $r_i \leftarrow R_{\text{uni}}(x_0^{(i)})$ for $i \in \{1..K\}$. ▷ *Eq. (9)*
  $A_i \leftarrow (r_i - \mu(r_{1:K}))/(\sigma(r_{1:K}) + \epsilon)$ ▷ *Per-group z-score*
  **% 3. On-Policy VLM Update (GRPO)**
  $\bar{\ell}_{\text{curr}} \leftarrow \frac{1}{|y|} \log \pi_\phi(y|\hat{x}_0^{(t)}, h_t), \quad \bar{\ell}_{\text{ref}} \leftarrow \frac{1}{|y|} \log \pi_{\text{sft_ref}}(y|\hat{x}_0^{(t)}, h_t)$
  $\widehat{\text{KL}} \leftarrow \bar{\ell}_{\text{curr}} - \bar{\ell}_{\text{ref}}$
  $\mathcal{L}_{\text{VLM}}(\phi) \leftarrow \frac{1}{N} \sum \left[ -A_i \cdot \bar{\ell}_{\text{curr}} + \beta_{\text{VLM}} \cdot \widehat{\text{KL}} \right]$
  $\phi \leftarrow \phi - \eta_{\text{VLM}} \nabla_\phi \mathcal{L}_{\text{VLM}}(\phi)$
  **% 4. Off-Policy DiT Update (DiffusionNFT)**
  **for** inner epoch $= 1$ **to** $3$ **do**
    Sample $\tau \in \{t\}_1^T$, $\epsilon \sim \mathcal{N}(0, I)$.
    $z_\tau \leftarrow (1 - \tau) z_0 + \tau \epsilon$
    $c_\tau \leftarrow \mathcal{C}[\tau]$ ▷ *Trajectory-Aware Replay*
    Compute implicit targets $v^{\pm}$ using $\beta_{\text{NFT}}$ and $v_{\text{old}}$.
    $\mathcal{L}_{\text{DiT}}(\theta) \leftarrow \mathbb{E}\left[ \frac{\tilde{r}}{\beta_{\text{NFT}}} \mathcal{L}^+ + \frac{1-\tilde{r}}{\beta_{\text{NFT}}} \mathcal{L}^- \right] \cdot A_{\max}$
    $\theta \leftarrow \theta - \eta_{\text{DiT}} \nabla_\theta \mathcal{L}_{\text{DiT}}(\theta)$
  **end for**
  **% 5. Hard Synchronization**
  $v_{\text{old}} \leftarrow v_\theta$
**end while**

# B CHAIN-OF-THOUGHT (COT) DATASET

To endow the student VLM (Bai et al., 2025) (Qwen2.5-VL-7B) with the capability to robustly estimate text content from degraded and partially restored images, we constructed a tailored Chain-of-Thought (CoT) dataset for Supervised Fine-Tuning (SFT). We base our dataset on the SA-Text training set (Min et al., 2025). To simulate authentic degradations, we apply the second-order degradation pipeline of Real-ESRGAN (Wang et al., 2021) (incorporating USM sharpening, a scale factor of 4, and LQ resizing) to generate low-quality (LQ) counterparts. Crucially, during the inference phase, the restoration model provides the VLM with progressively refined intermediate images and textual hints. To accurately mirror this dynamic environment during training, we include both high-quality (HQ) and LQ images in our dataset.

Data generation is driven by a teacher model (Qwen2.5-VL-32B), guided by a meticulously designed system prompt (see Figure 4). Rather than merely enforcing a structural `<think>-<answer>` format, the prompt instills a rigorous reasoning scheme. It enforces core principles such as "Visual Evidence First", "Anti-Hallucination", and a "Strict Latin Only" constraint, alongside word-level output separation. Furthermore, it dynamically modulates the depth of reasoning required based on the visual clarity of the text (i.e., clear versus blurry text). To simulate the diverse textual hint conditions encountered during inference, the user prompts are categorized into three distinct variants:

**Example of CoT Dataset Generation**

**System Prompt:**

You are a highly precise visual text reader providing OCR hints for image restoration.
Your goal is to read visible text accurately and safely infer missing parts based on context.

**Rules:**

1. **Visual Evidence First:** Always prioritize the exact characters and strokes you can clearly see.
2. **Safe Contextual Guessing:** If a word is truncated or blurred, recommend a word ONLY IF the visual background strongly supports it AND it perfectly matches the visible letters.
3. **Anti-Hallucination:** DO NOT make wild guesses. Output only what is strictly supported by visual evidence.
4. **Strict Language Constraint:** Output ONLY Latin letters, numbers, and basic punctuation. Non-Latin scripts (Chinese, Japanese, Korean, Arabic, Cyrillic, etc.) are STRICTLY PROHIBITED. If you see them, completely ignore them.
5. **Word-Level Output:** Output each word as a SEPARATE item in the JSON array. Do NOT merge multiple words into a single string with spaces. Example: `["NO", "SMOKING"]` NOT `["NO SMOKING"]`.

You MUST always use `<think>` tags before answering. Adjust your reasoning depth based on clarity:

*If the text is clear and easy to read:*
`<think>`
Text is clearly visible. Confident reading.
`</think>`
`<answer>`{"texts": ["WORD1", "WORD2"]}`</answer>`

*If the text is blurry, truncated, or ambiguous:*
`<think>`
- Trigger: [What part of the image is unclear]
- Visible Evidence: [What letters/strokes are clearly visible]
- Plausibility Check: [Does context justify a guess? Otherwise "Strict reading only"]
- Conclusion: [Final text, split into individual words, Latin only]
`</think>`
`<answer>`{"texts": ["WORD1", "WORD2"]}`</answer>`

**User Prompt (With Hints):**
A text detection model suggests: ["WORD1", "WORD2"].
Please verify and correct each word based on the image.

**User Prompt (No Hints):**
Please extract the visible text from this image.

Figure 4: The system and user prompts designed for generating the CoT dataset.

- **TESTR (Zhang et al., 2022)-hint:** Incorporates predictions from a pretrained text spotting model (TESTR). The detected words are sorted following a polygon-based reading order (top-to-bottom, left-to-right) and truncated to a maximum of 10 hints.
- **No-hint:** Applied when the TESTR model fails to detect any text, or employed intentionally as a regularization strategy to prevent the VLM from overly relying on explicit textual hints.
- **Corrupted-hint:** Designed to mimic realistic Optical Character Recognition (OCR) errors. We inject plausible corruptions into the Ground Truth (GT) text utilizing a visual confusion table (e.g., Y↔V, I↔l↔1, O↔0, rn↔m, ff↔fl). This perturbation is applied with a 40% probability per word, altering a maximum of two characters per word.

To ensure the high quality of the generated CoT rationales, we impose a stringent filtering protocol on the raw outputs from the teacher model. A sample is retained only if it satisfies all of the following criteria: (1) strict adherence to the `<think>-<answer>` format with a reasoning block of at least




**User prompt:**
Please extract the visible text from this image.

**VLM output:**
<think>
Text is clearly visible. Confident reading.
</think>
<answer>{"texts": ["JUVENTUS"]}</answer>




**User prompt:**
Please extract the visible text from this image.

**VLM output:**
<think>
Text is clearly visible. Confident reading.
</think>
<answer>{"texts": ["BTP"]}</answer>

Figure 5: Examples from our CoT dataset under the **no-hint** condition. For clearly visible text, the model performs direct, confident reading without requiring extensive step-by-step reasoning.

10 characters; (2) successful JSON parsing of the <answer> block; (3) exact set equality between the extracted text and the GT word set (penalizing both omissions and hallucinations); (4) exclusive use of Latin/ASCII characters; (5) length constraints to prevent infinite loops, bounded by 2,000 total characters and (6) 512 assistant tokens; and (7) a global cap limiting empty text predictions to a maximum of 5% of the dataset. Following this rigorous curation process, our final CoT dataset comprises 161,139 high-quality instruction pairs (approximately 120K HQ and 41K LQ samples). Illustrative examples of the dataset and prompts are provided in Fig. 5–7.

# C Additional Implementation Details

## C.1 Datasets and Evaluation Protocol

Following TeReDiff (Min et al., 2025), we utilize its benchmark datasets for both training (SFT and GRPO) and evaluation. Our model is trained on the SA-Text train set (100K $512 \times 512$ images), with synthetic degradations applied on-the-fly via the Real-ESRGAN pipeline (Wang et al., 2021). To comprehensively assess robustness, we evaluate on the SA-Text test set, which provides three progressive levels of synthetic degradation. Furthermore, to validate real-world generalization, we test on Real-Text, a collection of in-the-wild HR-LR pairs sourced from RealSR (Cai et al., 2019) and DRealSR (Pengxu Wei, 2020). Representative examples of both datasets are illustrated in Fig. 8.

## C.2 CoT Dataset Generation and Supervised Fine-Tuning

A Chain-of-Thought dataset consisting of approximately 161k instances was generated leveraging the Qwen2.5-VL-32B-Instruct model (Bai et al., 2025). Supervised Fine-Tuning (SFT) was subsequently performed on the Qwen2.5-VL-7B-Instruct (Bai et al., 2025) architecture using LoRA (Hu et al., 2022) ($r = 16$, $\alpha = 64$). The training was executed for 4 epochs optimized by AdamW (lr=$2 \times 10^{-5}$, weight decay=0.01), modulated by a cosine learning rate schedule with a 200 step linear warmup.




**User prompt:**
Please extract the visible text from this image.

**VLM output:**
<think>
- Trigger: The text is slightly blurry, but the letters are distinguishable.
- Visible Evidence: The letters "EDISON" are clearly visible on the building.
- Plausibility Check: The context of a building name supports the reading of "EDISON"
- Conclusion: The text is "EDISON"
</think>

<answer>{"texts": ["EDISON"]}</answer>




**User prompt:**
Please extract the visible text from this image.

**VLM output:**
<think>
- Trigger: The text is partially obscured and blurry, especially at the edges.
- Visible Evidence: The letters "LACOSTE" are clearly visible in the center of the image.
- Plausibility Check: The visible letters form a recognizable brand name, and there is no ambiguity in the visible portion. The edges are blurred, but the core text is distinct.
- Conclusion: Based on the visible evidence, the text is "LACOSTE."
</think>

<answer>{"texts": ["LACOSTE"]}</answer>

Figure 6: Examples from our CoT dataset under the **no-hint** condition featuring degraded text. When encountering blurry or partially obscured characters, the model automatically triggers a deeper, structured reasoning process. It carefully evaluates partial visual evidence and contextual plausibility to accurately deduce the text without hallucination.

### C.3 Asymmetric Joint RL Training

For the Asymmetric Joint RL phase, the VLM and DiT (Kim et al., 2025) are equipped with LoRA adapters ($r = 8$, $\alpha = 16$ and $r = 32$, $\alpha = 64$, respectively). Rollouts use a deterministic DPM-Solver-2 ODE ($T = 4$). We generate $K = 8$ stochastic VLM trajectories (temperature 0.3, top-p 0.95) for 2 unique images to form a batch of 16. The shared reward, evaluated on the final image via TESTR (Zhang et al., 2022), is $0.6 \cdot \text{det_F1} + 0.2 \cdot \text{e2e_none} + 0.2 \cdot \text{e2e_full}$. Advantages are computed via per-group z-score within the $K = 8$ candidates and clipped at 5.0. Both models are optimized with a learning rate of $1 \times 10^{-7}$ and hard synchronization (EMA 0.0) on a single H200 GPU. The DiT is updated off-policy via DiffusionNFT (3 inner epochs, $\beta_{\text{NFT}} = 0.1$, no KL). Concurrently, the VLM is updated on-policy via GRPO, applying a per-token KL penalty ($\beta_{\text{VLM}} = 0.001$) anchored to the frozen SFT adapter to maintain format compliance.

### C.4 Additional Optimization Details

**Asymmetric policy optimization.** Our Asymmetric Joint RL framework distinguishes between the environment's reverse transition density and the VLM's discrete autoregressive likelihood. Under the deterministic DPM-Solver-2, the reverse transition kernel is a Dirac measure with an undefined score, making standard likelihood-ratio policy gradients inapplicable to the DiT. In contrast, GRPO only requires the VLM's autoregressive likelihood of its generated textual actions, which can be computed through a teacher-forced forward pass. This distinction enables the VLM to be optimized on-policy via GRPO while the DiT is updated off-policy via DiffusionNFT.




**User prompt:**
A text detection model suggests: [”Baldc”].
Please verify and correct each word based on the image.

**VLM output:**
<think>
- Trigger: The text appears slightly blurred and the last letter is not fully clear.
- Visible Evidence: The visible letters are ”Balti”, with the last letter appearing to be a ”c” or ”e”.
- Plausibility Check: Given the context of the image, which resembles a map or geographical representation, ”Baltic” is a plausible word, referring to the Baltic region. The visible letters ”Balti” strongly support this guess.
- Conclusion: Based on the visible evidence and context, the most plausible word is ”Baltic”.
</think>

<answer>{”texts”: [”Baltic”]}</answer>




**User prompt:**
A text detection model suggests: [”BARRA”, ”BARRA”, ”OFF”, ”BEACH”, ”KEEP”, ”THE”, ”WHEN”, ”WIND3OCKS”, ”ARD”, ”FLYING”].
Please verify and correct each word based on the image.

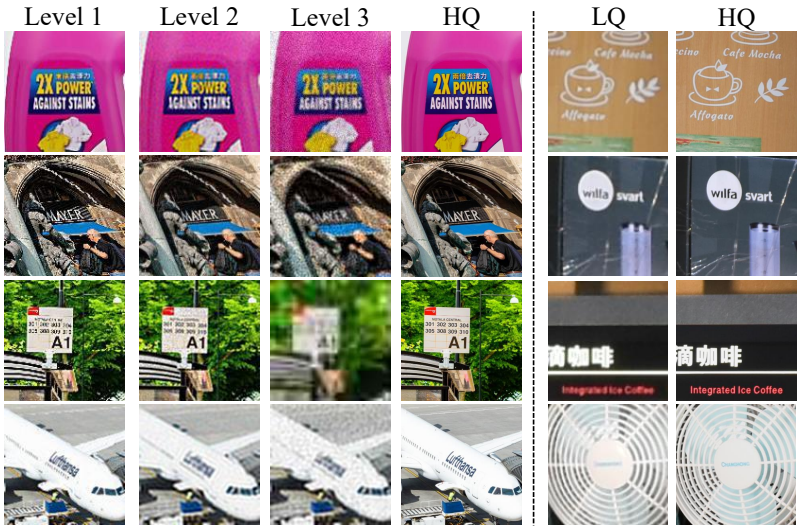
**VLM output:**
<think>
- Trigger: The text detection model has some inaccuracies, particularly in the words ”WIND3OCKS” and ”ARD.” These do not match the visible text in the image.
- Visible Evidence: ”BARRA” is clearly visible. ”AIRPORT” is visible below ”BARRA”. ”KEEP OFF THE BEACH” is clearly visible. ”WHEN WINDSOCKS ARE FLYING” is visible, with ”WINDSOCKS” and ”ARE” being the words that need correction.
- Plausibility Check: The context of the sign is related to airport safety, and the words ”WINDSOCKS” and ”ARE” are consistent with this context. The model’s suggestion of ”WIND3OCKS” and ”ARD” is incorrect.
- Conclusion: Based on the visible evidence and context, the corrected words are ”WINDSOCKS” and ”ARE”.
</think>

<answer>{”texts”: [”BARRA”, ”AIRPORT”, ”KEEP”, ”OFF”, ”THE”, ”BEACH”, ”WHEN”, ”WINDSOCKS”, ”ARE”, ”FLYING”]}</answer>

Figure 7: Examples from our CoT dataset under the **with-hint** condition for degraded text. When provided with an external text proposal, the model does not blindly trust the hint; instead, it critically cross-references the prior with partial visual evidence to deduce the accurate text.

**Per-step GRPO with trajectory-level advantages.** GRPO is applied at each denoising step using a shared trajectory-level advantage. Specifically, each of the $T = 4$ denoising steps triggers an independent VLM call conditioned on the intermediate state $s_t = (\hat{x}_0^{(t)}, h_t)$, generating a complete reasoning and answer sequence. The token log-likelihood of each sequence is evaluated individually, while the advantage is computed once per trajectory from the terminal reward across $K = 8$ candidates and shared across all four steps.

The dynamically changing textual condition $c_t$ requires no special handling in the VLM objective, as it represents the generated action rather than an external condition. Time variation enters through the intermediate visual state, to which we apply stop-gradient. Since reusing previous rollouts would




(a) SA-Text (b) Real-Text

Figure 8: **Representative examples from the evaluation datasets.** (a) SA-Text test set with three progressive levels of synthetic degradation and the corresponding HQ reference images. (b) Real-Text test set with representative real-world LQ-HQ image pairs.

introduce a mismatch between the stored states and the updated DiT, the VLM is optimized strictly on-policy with one gradient step per rollout.

### C.5 Computational Cost

Our full training pipeline takes approximately 44 hours on a single H200 GPU, comprising approximately 30 hours for CoT-format SFT and 14 hours for the complete Joint RL stage. The Joint RL time is measured end-to-end and includes trajectory rollout, reward computation, and both VLM and DiT optimization.

At inference time, our method invokes the VLM multiple times during restoration. The reported latency of 4.62 seconds per image includes all VLM calls and diffusion inference. Despite this additional autoregressive reasoning overhead, the latency remains comparable to UniT at 4.95 seconds per image. Our method uses vLLM for autoregressive VLM decoding. Since vLLM increased UniT's end-to-end latency, we report the faster official UniT implementation, while TeReDiff is also evaluated using its official implementation. All methods are benchmarked on the same hardware, with the same image resolution and batch size.

## D Further Analysis of Our Method

To further validate the effectiveness of our proposed framework, we extend the ablation study presented in Table 4 of the main paper with additional experiments. These analyses provide a more comprehensive understanding of the individual contributions of our core components and their interactions.

### D.1 Additional Ablation of CoT-format Supervision

To disentangle the effect of CoT-format supervision from domain-specific VLM fine-tuning, we introduce an additional answer-only (AO) SFT baseline, denoted as Setting B-AO. Settings B-AO and B share the same training data, model architecture, LoRA configuration, optimization hyperparameters, instruction rules, and structured answer schema, differing only in whether intermediate reasoning traces are supervised. Specifically, B-AO directly generates the `<answer>...</answer>` output, whereas B is additionally supervised with an intermediate `<think>...</think>` trace.

Table 5: **Effectiveness of CoT-format SFT.** End-to-end text spotting under the same restoration pipeline and four-step solver, extending the main ablation with the answer-only SFT variant.

| | | | ABCNet v2 | | | TESTR | | |
|---|---|---|---|---|---|---|---|---|
| Setting | SFT | CoT Trace | None (↑) | Full (↑) | Avg (↑) | None (↑) | Full (↑) | Avg (↑) |
| A | | | 30.96 | 39.93 | 35.45 | 31.99 | 41.52 | 36.76 |
| B-AO | ✓ | | 31.60 | 39.93 | 35.77 | 33.85 | 42.25 | 38.05 |
| B | ✓ | ✓ | **32.09** | **40.87** | **36.48** | **34.48** | **42.91** | **38.70** |

Table 6: **VLM word-level reading F1 (%).** Each model is evaluated at its best validation checkpoint.

| Model | Real-Text | SA-lv1 | SA-lv2 | SA-lv3 | 4-set Avg |
|---|---|---|---|---|---|
| B-AO (Answer-only SFT) | 56.08 | 40.85 | 38.08 | 29.11 | 41.03 |
| B (CoT-format SFT) | **60.96** | **42.24** | **38.56** | **30.18** | **42.99** |
| $\Delta$ (CoT contribution) | +4.88 | +1.39 | +0.48 | +1.07 | +1.96 |

Both models are evaluated at their respective best validation checkpoints using the same validation protocol.

As shown in Table 5, this comparison decomposes the A-to-B improvement into two components. Answer-only domain-specific SFT (A to B-AO) improves the E2E average scores by 0.32 and 1.29 points on ABCNet and TESTR, respectively. Adding intermediate reasoning-trace supervision (B-AO to B) provides further improvements of 0.71 and 0.65 points. Setting B consistently outperforms B-AO across both text spotters and all evaluation conditions, demonstrating the additional benefit of CoT-format supervision beyond domain-specific fine-tuning.

To further assess the effect of CoT-format supervision on VLM transcription, we evaluate the word-level reading F1 of the VLM output. As shown in Table 6, Setting B outperforms B-AO across all four evaluation sets, improving the average word-level F1 from 41.03 to 42.99. The largest improvement is observed on Real-Text, where the gain reaches 4.88 points. Together, the end-to-end and VLM-level results provide controlled empirical evidence that CoT-format supervision contributes beyond answer-only VLM fine-tuning.

### D.2 Sequential versus Joint Optimization

To further examine the benefit of joint co-evolution, we compare our framework with a sequential optimization strategy that first applies RL to the VLM, freezes it, and subsequently optimizes the DiT. Both strategies use the same SFT initialization, reward function, learning rate, and training configuration. Table 7 presents the results on SA-Text Level 1.

Our joint co-evolution framework outperforms sequential optimization across all six metrics, improving the E2E average scores from 39.10 to 39.44 on ABCNet and from 41.50 to 41.83 on TESTR.

During the VLM-only RL stage, the frozen renderer limits the benefit of optimizing textual guidance, as the VLM receives reward only through what the fixed DiT can render. Once the DiT is subsequently updated, the frozen VLM cannot adapt its guidance policy to the evolving intermediate reconstructions. In contrast, joint co-evolution allows both models to adapt concurrently under a shared reward. These results provide additional empirical support for concurrent optimization, although the observed gains over sequential training are modest.

### D.3 Effect of Denoising Steps

To examine whether Progressive Search remains effective beyond the default four-step trajectory, we conduct an additional ablation over the number of ODE steps $T$. Table 8 reports the E2E recognition performance on SA-Text Level 1.

Table 7: **Sequential versus joint optimization.** E2E recognition F1 (%) on SA-Text Level 1. Bold indicates the higher score.

| | ABCNet v2 | | | TESTR | | |
|---|---|---|---|---|---|---|
| Method | None (↑) | Full (↑) | Avg (↑) | None (↑) | Full (↑) | Avg (↑) |
| Sequential (VLM-RL → freeze → DiT-RL) | 35.21 | 42.98 | 39.10 | 37.57 | 45.42 | 41.50 |
| Joint co-evolution (Ours) | **35.58** | **43.46** | **39.52** | **37.82** | **45.47** | **41.65** |

Table 8: **Effect of denoising steps.** E2E recognition F1 (%) on SA-Text Level 1 with different numbers of ODE steps. Bold indicates the best result.

| | ABCNet v2 | | | TESTR | | |
|---|---|---|---|---|---|---|
| Steps | None (↑) | Full (↑) | Avg (↑) | None (↑) | Full (↑) | Avg (↑) |
| $T = 4$ | 35.58 | 43.46 | 39.52 | 37.82 | 45.47 | 41.65 |
| $T = 6$ | 37.89 | 46.11 | 42.00 | 38.20 | 46.52 | 42.36 |
| $T = 8$ | 37.95 | 46.16 | 42.05 | 38.71 | 47.06 | 42.89 |

Increasing $T$ from 4 to 8 improves all six metrics, with gains of 2.61 and 1.06 points in ABCNet and TESTR E2E average scores, respectively. Longer denoising trajectories provide additional intermediate observations, allowing the VLM further opportunities to refine its textual hypotheses as $\hat{x}_0^{(t)}$ becomes clearer.

We adopt $T = 4$ as the default configuration to balance restoration quality and computational cost, since each additional step requires an extra VLM inference and DiT solver call. Increasing the number of inference steps offers a practical accuracy–latency trade-off.

### D.4 Early-Stage Prediction Errors and Progressive Refinement

To examine the sensitivity of the final output to the initial text prediction, we freeze the text condition predicted at the first denoising step and reuse it unchanged throughout all remaining steps. This disables subsequent inner-loop refinement, preventing errors or incomplete predictions in the initial text condition from being corrected. We apply this intervention separately to Settings D and F while keeping the architecture, weights, and denoising trajectory unchanged.

As shown in Table 9, freezing the first-step condition reduces the E2E average scores of our full model by 1.19 and 1.63 points on ABCNet and TESTR, respectively. The performance reduction is larger for Setting F than for Setting D, indicating that our full framework benefits more from subsequent intermediate observations to revise and complete its textual guidance.

Unlike existing methods that do not expose comparable multi-step text-prediction trajectories, our frozen-condition variant provides a controlled single-shot baseline within the same framework. These results demonstrate the benefit of progressive re-reading in mitigating the effects of incomplete or inaccurate initial predictions.

### D.5 Multi-seed Evaluation of Joint Co-evolution

To further validate the effectiveness of joint co-evolution, we evaluate DiT-only RL (Setting D) and our full framework (Setting F) across three independent random seeds. Table 10 reports the mean and standard deviation of E2E recognition F1 on SA-Text Level 1, evaluated using ABCNet v2 and TESTR.

Across three independent random seeds, our full co-evolutionary framework achieves higher mean performance than Setting D across all six evaluated metrics, including ABCNet E2E-Full (43.42 vs. 43.39). Notably, the slight E2E-Full inversion observed in the single-run ablation (43.29 vs. 43.54) does not persist in the multi-seed averages.

Joint co-evolution improves the mean E2E average scores from 39.33 to 39.44 on ABCNet and from 41.16 to 41.37 on TESTR. Although the gains are modest relative to the observed variation,

Table 9: **Effectiveness of inner-loop refinement.** E2E recognition F1 (%) on SA-Text Level 1. Fixed denotes reusing the initial text condition throughout denoising. Values in parentheses indicate improvements from refinement.

| | ABCNet v2 | | | TESTR | | |
|---|---|---|---|---|---|---|
| Setting | None (↑) | Full (↑) | Avg (↑) | None (↑) | Full (↑) | Avg (↑) |
| D w/o inner-loop reasoning | 34.81 | 42.25 | 38.53 | 36.86 | 44.36 | 40.61 |
| D | 35.12 **(+0.31)** | 43.54 **(+1.29)** | 39.33 **(+0.80)** | 37.04 **(+0.18)** | 45.12 **(+0.76)** | 41.08 **(+0.47)** |
| F w/o inner-loop reasoning | 34.64 | 41.85 | 38.25 | 36.65 | 43.74 | 40.20 |
| F | 35.58 **(+0.94)** | 43.46 **(+1.61)** | 39.52 **(+1.27)** | 37.82 **(+1.17)** | 45.47 **(+1.73)** | 41.65 **(+1.45)** |

Table 10: **Multi-seed evaluation of joint co-evolution.** E2E recognition F1 (%) on SA-Text Level 1, reported as mean ± standard deviation over three seeds. Bold indicates the higher mean.

| | ABCNet v2 | | | TESTR | | |
|---|---|---|---|---|---|---|
| Method | None (↑) | Full (↑) | Avg (↑) | None (↑) | Full (↑) | Avg (↑) |
| DiT-only RL (D) | $35.27 \pm 0.16$ | $43.39 \pm 0.18$ | $39.33 \pm 0.03$ | $37.24 \pm 0.34$ | $45.08 \pm 0.05$ | $41.16 \pm 0.18$ |
| Joint RL (F) | $\mathbf{35.45 \pm 0.55}$ | $\mathbf{43.42 \pm 0.20}$ | $\mathbf{39.44 \pm 0.19}$ | $\mathbf{37.42 \pm 0.46}$ | $\mathbf{45.32 \pm 0.44}$ | $\mathbf{41.37 \pm 0.43}$ |

the positive mean improvements across both the reward-based TESTR evaluator and the held-out ABCNet evaluator provide additional evidence of the benefits of concurrent co-evolution.

### D.6 Cross-Evaluator Generalization

To examine whether our improvements are specific to the TESTR-based reward, we further analyze performance using the held-out ABCNet v2 evaluator. Notably, both UniT and TeReDiff incorporate TESTR during training: TeReDiff uses TESTR to generate textual prompts, while UniT trains TESTR on diffusion features to obtain intermediate OCR predictions. Our evaluation therefore follows the established benchmark protocol while additionally reporting results from a separate text-spotting architecture.

Our method achieves consistent improvements on ABCNet v2, which is not used for reward computation. On the Real-Text benchmark, our framework achieves an ABCNet E2E average score of 55.64%, outperforming UniT (52.96%). Furthermore, on SA-Text Level 1, our full model improves ABCNet E2E-None from 30.96% to 35.58% over the baseline (Setting A), yielding a gain of 4.62 points without lexicon assistance. These cross-evaluator improvements provide additional evidence that our framework enhances text legibility beyond the specific evaluator used for reward optimization.

### D.7 Optimization Stability

To assess the stability of our Asymmetric Joint RL framework, we investigate the effects of optimization hyperparameters and monitor training behavior using held-out evaluation metrics. In preliminary experiments, larger learning rates of $10^{-5}$ and $10^{-6}$ caused the VLM's structured output to degrade during prolonged training, while excessive weighting of the text-spotting reward introduced mild over-sharpening artifacts.

Based on these observations, we adopt a learning rate of $10^{-7}$ for both models, limiting per-update policy shifts and allowing hard synchronization (EMA 0.0). The DiT update is further stabilized through implicit positive/negative mixing ($\beta_{\mathrm{NFT}} = 0.1$), a bounded soft positive ratio $\tilde{r} \in [0, 1]$, and advantage clipping at 5.0. Meanwhile, the VLM is regularized by a KL penalty anchored to the frozen SFT policy and a format gate to preserve structured output compliance.

We further monitor training stability using periodic held-out evaluations on Real-Text, tracking the VLM's intrinsic word-level F1, external text-spotting detection and E2E scores, and generated reasoning traces. Under the final configuration, no predefined instability threshold was triggered, and the format-gate pass rate remained approximately 1.0 throughout training. We observed no sustained

Table 11: **Robustness to stochastic sampling.** E2E recognition F1 (%) on SA-Text Level 1. Stochastic results are reported as mean ± standard deviation over three random seeds. Bold indicates the higher mean score.

| | ABCNet v2 | | | TESTR | | |
|---|---|---|---|---|---|---|
| Sampling | None (↑) | Full (↑) | Avg (↑) | None (↑) | Full (↑) | Avg (↑) |
| Deterministic | 35.58 | 43.46 | 39.52 | 37.82 | 45.47 | 41.65 |
| Stochastic ($\eta = 0.1$, 3 seeds) | $\mathbf{37.76 \pm 0.37}$ | $\mathbf{45.50 \pm 0.58}$ | $\mathbf{41.63 \pm 0.47}$ | $\mathbf{38.74 \pm 0.46}$ | $\mathbf{46.75 \pm 0.34}$ | $\mathbf{42.75 \pm 0.32}$ |

Table 12: **Perceptual quality comparison** averaged over SA-Text Levels 1–3 and Real-Text. We report the full-reference metric LPIPS and no-reference metrics NIQE, MUSIQ, CLIP-IQA, MANIQA, and BRISQUE. Bold indicates the best result among the matched four-step methods. The HQ row reports the no-reference metric statistics of ground-truth images for reference.

| Method | LPIPS (↓) | NIQE (↓) | MUSIQ (↑) | CLIP-IQA (↑) | MANIQA (↑) | BRISQUE (↓) |
|---|---|---|---|---|---|---|
| UniT | 0.355 | 10.81 | 62.70 | 0.562 | 0.513 | 64.27 |
| TeReDiff | **0.280** | 10.26 | 64.97 | 0.588 | **0.565** | 38.53 |
| Ours | 0.320 | **7.59** | **67.17** | **0.615** | 0.522 | **34.07** |
| HQ (ref.) | – | 5.21 | 65.61 | 0.583 | 0.482 | 28.66 |

divergence, model collapse, or visual artifacts, demonstrating stable optimization under the adopted training configuration.

### D.8 Robustness to Stochastic Sampling

To examine whether our framework remains effective under stochastic sampling, we replace the deterministic DPM-Solver-2 with a stochastic SDE sampler following the Flow-GRPO formulation ($\eta = 0.1$). We keep the model checkpoint, denoising steps ($T = 4$), and greedy VLM decoding unchanged. Both sampling strategies are evaluated on the full SA-Text Level 1 test set (1,000 images), with stochastic results averaged over three random seeds.

As shown in Table 11, stochastic sampling improves all six E2E recognition metrics, increasing the ABCNet and TESTR average scores from 39.52 to 41.63 and from 41.65 to 42.75, respectively, with low variation across seeds.

The improvement is accompanied by higher detection precision and lower recall, suggesting that mild injected noise may suppress marginal, barely legible glyphs that are detected but misrecognized under deterministic sampling. Furthermore, stochastic sampling allows partial revision of glyph errors introduced during early denoising, complementing our Progressive Search mechanism. These results demonstrate that our framework maintains competitive restoration performance under stochastic inference, beyond its default deterministic sampling configuration.

### D.9 Perceptual Quality Analysis

To examine whether improvements in text legibility are accompanied by a degradation in overall image quality, we evaluate complementary full-reference and no-reference perceptual metrics. Table 12 presents the results averaged over SA-Text Levels 1–3 and Real-Text, comparing methods under the same four-step diffusion budget. We additionally report the no-reference metric values of the HQ ground-truth images for reference.

**Comparison with existing methods.** Our method achieves favorable no-reference perceptual quality, outperforming competing approaches on NIQE, MUSIQ, CLIP-IQA, and BRISQUE. While TeReDiff obtains better LPIPS and MANIQA scores, our method maintains competitive perceptual quality while significantly improving text legibility. These results demonstrate that our text-oriented optimization preserves overall visual quality under the same diffusion-step budget.

**Perceptual quality relative to HQ images.** Our outputs achieve no-reference scores comparable to those of HQ images across several metrics. In particular, our method reaches 67.17 versus 65.61 on MUSIQ, 0.615 versus 0.583 on CLIP-IQA, and 0.522 versus 0.482 on MANIQA. Since no-

reference metrics may favor properties such as sharpness, scores exceeding the HQ mean do not necessarily indicate higher fidelity than the ground truth. Nevertheless, these results suggest that the improvements in functional legibility are achieved without a substantial degradation in overall perceptual quality.

## E LIMITATIONS

This work has several limitations. First, under extreme degradation with zero visual evidence, the VLM may rely on language priors and generate plausible but incorrect hallucinations. To mitigate this, future work should incorporate confidence-aware abstention to safely handle fundamentally irrecoverable text. Second, the framework faces a capacity bottleneck in highly text-dense scenes because external text hints are truncated to a maximum of 10 words. While necessary for structured reasoning, this limits efficacy on crowded images and requires future scaling to process dense typography effectively.

## F BROADER IMPACT

This work introduces a novel framework for text-aware image restoration by enabling the joint evolution of Vision-Language Models (VLMs) and Diffusion Transformers. While the primary focus is on fundamental algorithmic improvements, we recognize several potential societal impacts:

**Positive Societal Impacts** The proposed method can significantly benefit digital archiving and the preservation of historical documents where text is often degraded or partially lost. By providing high-fidelity text restoration, it improves the accessibility of information for researchers and the general public. Furthermore, the progressive visual reasoning capability can be integrated into assistive technologies for the visually impaired, helping them read signs or documents in challenging real-world environments more accurately.

**Negative Societal Impacts and Mitigations** As with any advanced generative model, there is a potential risk that high-fidelity text restoration could be misused for malicious purposes, such as document forgery or the manipulation of visual evidence (e.g., altering license plates or legal documents). To mitigate these risks, we emphasize that our model is designed for restoration and accessibility. We encourage the community to use this technology alongside digital forensics and watermarking techniques to verify the authenticity of restored images. Additionally, our reliance on open-source datasets and models ensures transparency, allowing the research community to better monitor and audit the behaviors of such generative frameworks.