FIXING THE LAST MILE: FAILURE-AWARE IMAGE RESTORATION WITH REGION-SPECIFIC GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in image restoration have greatly improved global image fidelity, producing visually plausible results across most regions of an image. Nevertheless, even high-quality restorations often exhibit localized failures, including artifacts in semantically important regions such as faces and text, structural inconsistencies, and texture mismatches. Although these errors typically affect only a small portion of the image, they often dominate the overall perceptual quality of the result. In this work, we propose **Failure-Aware Image Restoration (FAIR)** to address the **"last-mile"** problem in image restoration. Rather than applying one uniform generative prior over the whole image, we condition a single-pass restoration model on both **where** local failures are likely to occur and **what** corrective guidance each region requires. This region-bound conditioning enables fine-grained and controllable restoration that improves semantic accuracy and stylistic consistency in critical regions while maintaining faithful, high-quality restoration elsewhere. Our method complements existing restoration pipelines by shifting the emphasis from global fidelity to local perceptual acceptability. Extensive experiments show that the proposed approach resolves common failure cases and improves perceptual quality, bridging the gap between technical accuracy and practical usability.

1 INTRODUCTION

Recent advances in perception-oriented image restoration (Lin et al., 2024; Yu et al., 2024; Wu et al., 2024b;a; Chen et al., 2025b; Lin et al., 2025) have substantially improved overall image quality, producing results that are globally plausible and perceptually rich. While existing methods have demonstrated impressive performance, we identify an overlooked class of **"last-mile"** problems in image restoration that fundamentally hinders their real-world deployment.

A near-perfect landscape with a distorted face, or a sharp document with garbled characters, is rarely considered acceptable regardless of how well the remaining regions are recovered. As shown in Fig. 1, such localized failures arise consistently under severe degradations, where heavy blur, noise, or compression remove local information beyond recovery, forcing generative priors to hallucinate plausible yet unfaithful content. Although these errors typically affect only a small portion of the image, they often dominate the perceived quality of the result. We refer to this challenge as the **"last-mile"** problem in image restoration: closing the perceptual gap by predicting failure-prone regions from the low-quality input and guiding their restoration with external references.

The last-mile problem stems from a structural limitation of current restoration pipelines. Resolving it requires two capabilities that existing methods lack: external information to recover content that is irretrievable from the degraded input, and fine-grained local control to apply such information only where it is needed. Yet existing approaches typically cast restoration as a global reconstruction problem, without explicitly modeling **where** local failures are likely to occur and, within each at-risk region, **what** type of failure it is prone to and what corrective guidance it requires. Consequently, the same generative prior is applied uniformly across the image, from smooth backgrounds to semantically important areas such as faces, text, and logos.

The challenge is to introduce external information selectively: it should help regions with insufficient recoverable content without unnecessarily altering regions well constrained by the input. Even recent semantic-guided methods (Yu et al., 2024; Wu et al., 2024b; Sun et al., 2024) introduce high-level priors at the image level, while deriving them largely from the corrupted input. When local

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.

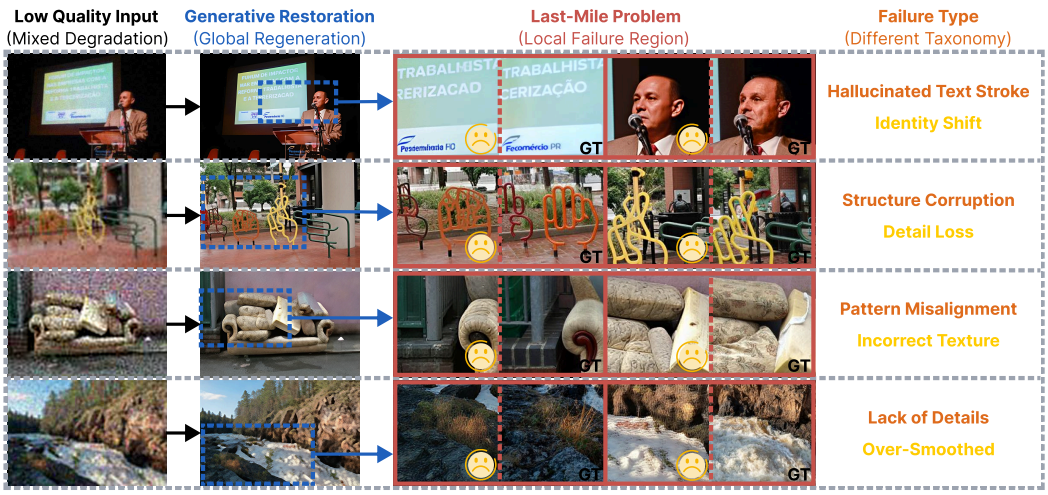

Figure 1: **The "last-mile" problem in generative image restoration.** State-of-the-art methods (Yu et al., 2024; Wu et al., 2024b; Lin et al., 2025) produce globally plausible reconstructions from severely degraded inputs, yet localized failures persist. Despite their small spatial extent, these failures dominate perceived quality and practical acceptability.

information is severely lost, these priors can generate plausible details without ensuring fidelity to the original content. Restoration therefore needs evidence beyond the input and spatial control to apply it only where needed. Deferring guidance to post-hoc repair has further costs: it may disturb well-restored regions, while the second model must work from an image already altered by the first, potentially missing original cues or containing new artifacts.

We propose **FAIR**, a failure-aware image restoration framework that analyzes the low-quality input, predicts failure-prone regions and their types, and prepares matching conditions before restoration. Rather than treating the image as a uniform field, FAIR recognizes that under severe degradation, small but perceptually critical regions often require different forms of corrective information. To study this behavior systematically, we construct a large-scale failure-aware dataset by applying state-of-the-art restoration models to degraded inputs and automatically annotating both **where** failures occur and **what** type of failure is present. Each failure region is additionally paired with an appropriate problem description, yielding region-level supervision aligned with our failure taxonomy. We observe a consistent pattern in this dataset: restoration failures cluster into a small set of semantically meaningful categories, each requiring different forms and strengths of external guidance.

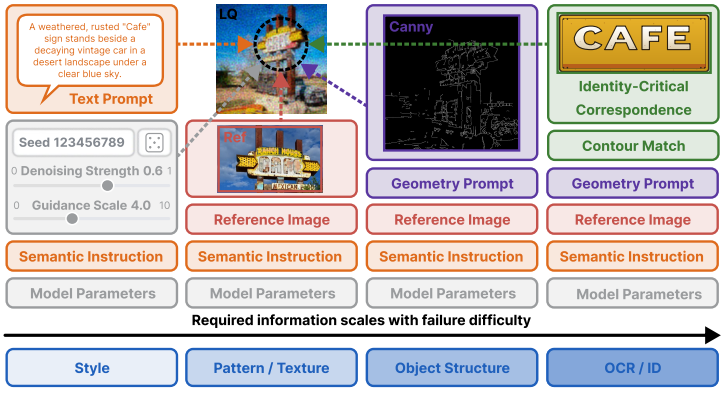

Figure 2: **Failure-aware guidance for local restoration challenges.** We group typical local restoration failures into four categories (bottom row) and show the corresponding condition combinations (top). As local content becomes harder to infer from the low-quality input, guidance progresses from semantic instructions and parameter controls to reference images, geometric contours, and identity correspondences. Region control limits where these conditions act.

As illustrated in Fig. 2, we organize restoration guidance by increasing constraint strength, with more specific guidance addressing progressively harder local failures: (i) semantic instructions and parametric controls for subjective artifacts; (ii) reference exemplars for pattern and texture inconsis-

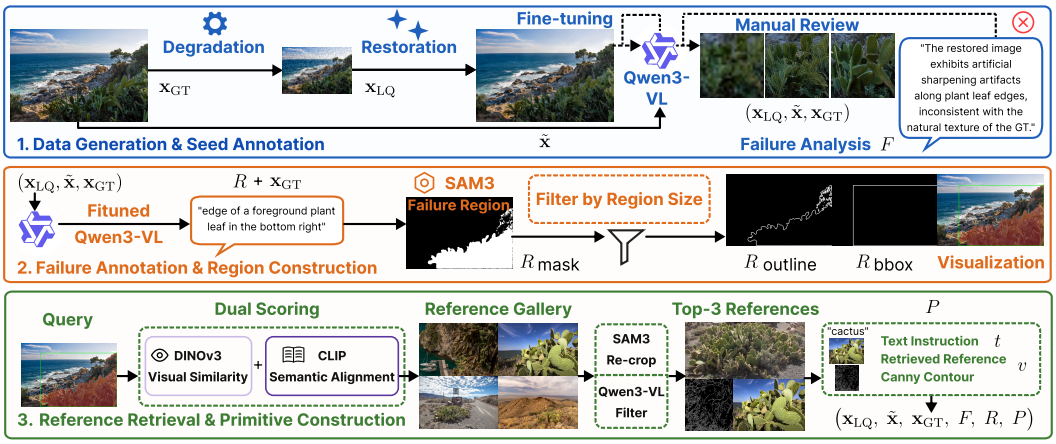

Figure 3: **Three-stage offline construction of failure-aware training data.** Stage I degrades $\mathbf{x}_{\mathrm{GT}}$ to $\mathbf{x}_{\mathrm{LQ}}$, produces a coarse restoration $\tilde{\mathbf{x}}$, and uses screened annotations to fine-tune Qwen3-VL-8B. Stage II supplies $(\mathbf{x}_{\mathrm{LQ}}, \tilde{\mathbf{x}}, \mathbf{x}_{\mathrm{GT}})$ to the fine-tuned model to identify failure type F and a localization phrase; SAM3 converts the phrase into region control R. Stage III retrieves and filters local references to build restoration guidance $P = (t, v)$, with text instruction t and optional visual guidance v.

tencies; (iii) geometric prompts for fine structural correction; and (iv) identity correspondences for identity-critical regions. Different failures call for different forms of guidance, and FAIR accepts each of them as a condition bound to the target region. Unlike existing methods that apply uniform generation across the image, FAIR binds each guidance condition to its target region. Built on FLUX.2-klein-9B (Labs, 2025) with lightweight LoRA modules, FAIR uses region control to specify **where** additional guidance is provided and a semantic reference to indicate **what** the restoration should follow. Training adopts a two-stage curriculum: single-step restoration pretraining, followed by region-anchored fine-tuning with external guidance, spatial control, and region-aware reconstruction reweighting. Experiments show that FAIR improves perceptual acceptability over strong baselines while maintaining high-quality restoration elsewhere.

2 Related Work

Image Restoration. Recent progress in image restoration has been largely driven by deep learning, leading to substantial improvements in super-resolution, denoising, and deblurring. Early CNN-based approaches (Dong et al., 2015; Lim et al., 2017; Zhang et al., 2018b; Dai et al., 2019; Mei et al., 2021) learn direct mappings from degraded inputs to clean images and achieve strong performance under synthetic degradations. Transformer-based methods (Liang et al., 2021; Wang et al., 2022; Zamir et al., 2022; Chen et al., 2023; Wu et al., 2025) further improve reconstruction quality through long-range dependency modeling and more expressive architectures. GAN-based approaches (Ledig et al., 2017; Wang et al., 2018; 2021; Liang et al., 2022) focus on perceptual realism by synthesizing high-frequency details through adversarial training. Despite these advances, most existing methods remain optimized for global reconstruction objectives such as pixel-level fidelity or perceptual similarity. As a result, they often fail in localized regions, particularly semantically important areas such as faces and text, where even minor artifacts can substantially degrade perceptual quality.

Diffusion-Based Restoration. Diffusion models (Ho et al., 2020; Song et al., 2020; Dhariwal & Nichol, 2021; Rombach et al., 2022) have become a dominant paradigm for image restoration (Lin et al., 2024; Yu et al., 2024; Wu et al., 2024a; Chen et al., 2025b; Lin et al., 2025; Saharia et al., 2022; Dong et al., 2025), achieving strong perceptual quality across a wide range of tasks. However, image-level conditioning alone often struggles under severe degradations. Recent studies report a shift from under-generation to over-generation, where strong generative priors produce hallucinations, structural distortions, and semantic inconsistencies in challenging regions such as faces, fine structures, and textured surfaces (Yin et al., 2026). To address these issues, several methods introduce explicit semantic guidance. PASD (Yang et al., 2024) extracts semantic cues from low-resolution inputs using off-the-shelf vision models and injects them through pixel-aware cross-attention. SeeSR (Wu et al., 2024b) generates degradation-aware tag prompts, while SUPIR (Yu et al., 2024) and CoSeR (Sun

et al., 2024) leverage multimodal LLMs to produce richer semantic descriptions. InstructRestore (Liu et al., 2025) further introduces user-instruction control, and CODSR (Chen et al., 2025a) reformulates restoration as one-step generation with region-adaptive priors. Nevertheless, these approaches remain primarily driven by global objectives and image-level priors, without explicitly predicting **where** restoration is likely to fail or what guidance each region needs.

Reference-Based Restoration. Reference-based restoration uses external high-quality images to recover details irretrievable from severely degraded inputs. Early methods rely on patch matching and texture transfer (Zhang et al., 2019; Yang et al., 2020; Lu et al., 2021; Jiang et al., 2021; Cao et al., 2022), typically assuming curated, strongly aligned references. Recent diffusion-based approaches relax this requirement: ReFIR (Guo et al., 2024) injects retrieved features through cross-image attention, RASR (Yan et al., 2025) formulates retrieval-augmented reference super-resolution, and AdaRefSR (Wang et al., 2026) adaptively weights reference features according to local relevance. Nevertheless, these methods use references as a primary condition for overall reconstruction without explicitly modeling failure-prone regions. FAIR instead confines external guidance to those regions. FAIR instead binds guidance to the relevant failure-prone region, confining external evidence there.

3 METHOD

Given a low-quality input $\mathbf{x}_{\mathrm{LQ}}$, we produce a restored image $\hat{\mathbf{x}}$ by applying dedicated guidance to failure-prone regions while restoring the remaining regions without additional guidance:

$$\hat{\mathbf{x}} = f_\theta\big(\mathbf{x}_{\mathrm{LQ}}, R, P\big). \tag{1}$$

Here, R specifies **where** to provide additional guidance using a localization phrase, bounding box, outline, or mask. The restoration guidance $P = (t, v)$ specifies **what** information to provide, combining a textual instruction t with optional visual guidance v from a retrieved reference or Canny contour; Appendix B maps the guidance types of Fig. 2 onto P. At inference, R and P are predicted from $\mathbf{x}_{\mathrm{LQ}}$ or provided by a user. Let $\mathcal{F} = \{F_1, F_2, F_3, F_4\}$ denote the failure tiers in Sec. 1; $F \in \mathcal{F}$ indicates which form of P a region typically needs. Binding P to R provides additional information to failure-prone regions during restoration. FAIR comprises a failure-aware dataset (Sec. 3.1), localized multimodal conditioning (Sec. 3.2.1), and a two-stage region-aware curriculum (Sec. 3.2.2).

3.1 FINE-GRAINED FAILURE-AWARE DATA CURATION

Training a failure-aware model requires supervision beyond standard $(\mathbf{x}_{\mathrm{LQ}}, \mathbf{x}_{\mathrm{GT}})$ pairs. For each pair, we generate a coarse restoration $\tilde{\mathbf{x}}$ and structured annotations (F, R, P) specifying the failure type, target region, and corresponding restoration guidance:

$$\big(\mathbf{x}_{\mathrm{LQ}},\ \tilde{\mathbf{x}},\ \mathbf{x}_{\mathrm{GT}},\ F,\ R,\ P\big). \tag{2}$$

As illustrated in Fig. 3, our failure-aware data construction pipeline proceeds in three stages.

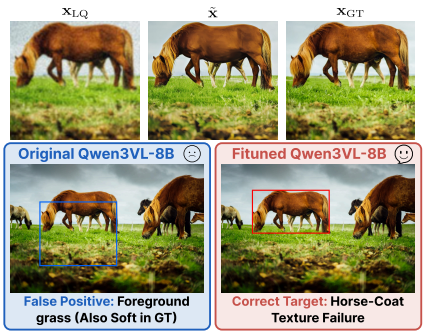

Figure 4: **Failure localization before and after fine-tuning.** The off-the-shelf model selects foreground grass (blue), while the fine-tuned model selects the central horse's mane and tail (red).

(1) Data generation and seed annotation. We draw high-quality images $\mathbf{x}_{\mathrm{GT}}$ from LSDIR and Unsplash (Li et al., 2023; Unsplash, 2020), with FFHQ (Karras et al., 2019) added to the restoration mix for face coverage, synthesize $\mathbf{x}_{\mathrm{LQ}}$ using the second-order Real-ESRGAN pipeline (Wang et al., 2021), and process it with multiple state-of-the-art restoration models to obtain coarse restorations $\tilde{\mathbf{x}}$ exhibiting realistic local failures. For each $(\mathbf{x}_{\mathrm{LQ}}, \tilde{\mathbf{x}}, \mathbf{x}_{\mathrm{GT}})$ triplet, Qwen3-VL-8B (Bai et al., 2025) predicts a failure label F and a phrase identifying the affected region. A structured prompt defines the failure taxonomy, output format, and localization requirements. Because the off-the-shelf model may mislocalize failures (Fig. 4), we manually verify its outputs against the ground truth and retain 1K triplet–annotation pairs with correct failure labels and localization. These verified pairs are used to fine-tune Qwen3-VL-8B with rank-64 LoRA, yielding the annotator used in Stage 2.

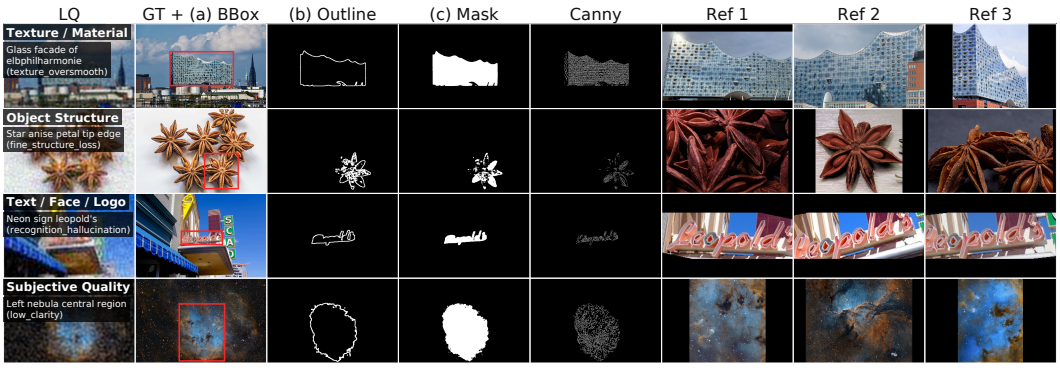

Figure 5: **Representative failure-aware dataset samples, one per tier.** Each shows an LQ–GT pair, four region controls (localization phrase, box, outline, mask), restoration guidance $P = (t, v)$ (textual instruction t with optional visual guidance v), and failure label F.

(2) Failure annotation and region construction. The fine-tuned annotator processes the full pool with the same input triplet, producing a failure label and localization phrase. SAM3 (Carion et al., 2025) segments the corresponding region in $\mathbf{x}_{\mathrm{GT}}$; after filtering masks with extreme area ratios, we derive a bounding box and closed outline. The phrase, box, outline, and mask form four interchangeable controls R; after the final Stage-3 filtering procedure described in Appendix B.4, 67K regions remain. At test time, condition prediction uses $\mathbf{x}_{\mathrm{LQ}}$ alone; its accuracy and GT-assisted oracle are evaluated in Appendix F.

(3) Reference retrieval and guidance construction. For each annotated region, we retrieve candidate patches from a multi-scale LSDIR–Unsplash gallery, ranked by DINOv3 (Siméoni et al., 2025) similarity to the target region and CLIP (Radford et al., 2021) consistency with its semantic description. SAM3 re-crops the corresponding instances, while Qwen3-VL-8B filters references inconsistent with the region description. We retain up to three cross-image references from distinct source images. The restoration guidance $P = (t, v)$ combines textual instruction t with visual guidance v sampled as empty, a visual reference, or a Canny contour. Appendix B.3 details the retrieval, filtering, and guidance-selection procedures.

Dataset summary. The resulting dataset contains 67K annotated failure regions spanning the four failure tiers. Each region is associated with a failure label F and multiple compatible forms of region control R and restoration guidance P. During training, one compatible (R, P) combination is randomly sampled for each region, exposing the model to diverse spatial and corrective conditions. Fig. 5 shows representative samples from each tier; additional examples, category distributions, and dataset statistics are provided in Appendix B.

3.2 Two-Stage Region-Guided Restoration Framework

FAIR adapts FLUX.2-klein-9B (Labs, 2025) for restoration conditioned on the low-quality image, region control R, and restoration guidance P:

$$\hat{\mathbf{x}} \sim p_\theta\big(\hat{\mathbf{x}} \,\big|\, \mathbf{x}_{\mathrm{LQ}}, R, P\big)$$

using a fixed-timestep flow-matching formulation (Lipman et al., 2022; Esser et al., 2024). The guidance $P = (t, v)$ is defined in Sec. 3.1, and R localizes its effect.

3.2.1 Architecture and Localized Guidance Conditioning

As illustrated in Fig. 6, FAIR jointly conditions restoration on the low-quality input $\mathbf{x}_{\mathrm{LQ}}$, region control R, and restoration guidance P: the instruction t enters the text stream, while spatial forms of R and visual guidance v enter the image stream, so that R localizes the guidance (Appendix C.2). Training follows two stages: the first establishes a general restoration prior, and the second introduces localized guidance and region-aware supervision.

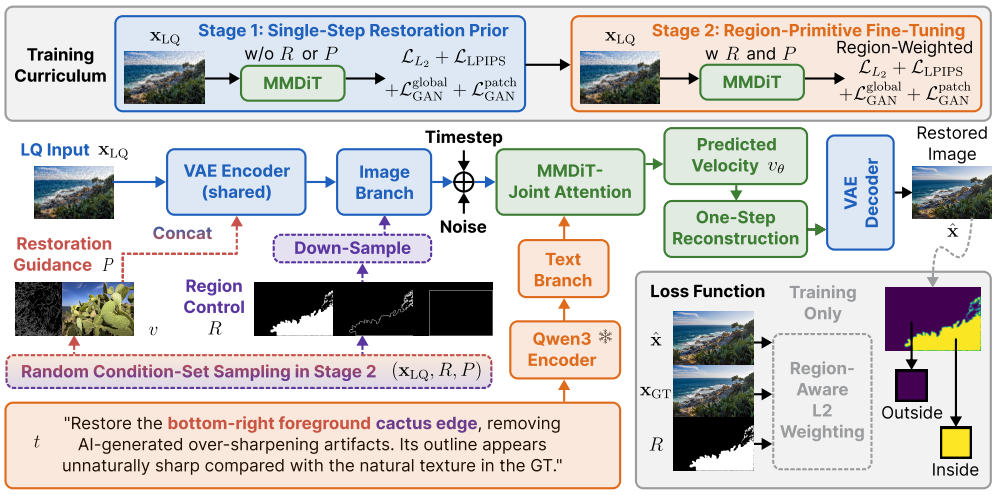

Figure 6: **Overview of FAIR's two-stage training.** Stage 1 learns a general restoration prior. Stage 2 conditions restoration on region control R and guidance $P = (t, v)$, using region-aware supervision to bind the guidance to its target region.

3.2.2 Two-stage curriculum

Stage 1: general restoration prior. We fine-tune the LoRA-adapted backbone on LQ–GT pairs at 1024^2 resolution. At a fixed timestep, the model predicts v_θ from noisy LQ latent z_t, reconstructs $z_0 = z_t - \sigma_t v_\theta$, and decodes it as $\hat{\mathbf{x}} = \mathcal{D}(z_0)$. The generator objective is

$$\mathcal{L}_G^{\text{S1}} = \lambda_{L_2}\mathcal{L}_{L_2} + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}} + \lambda_{\text{GAN}}\mathcal{L}_{\text{GAN}}, \tag{3}$$

All three losses act on the decoded image. The GAN loss uses an R1-regularized ConvNeXt discriminator. Loss weights and regularization settings are given in Appendix C.

Stage 2: region-guided fine-tuning. We fine-tune the shared LoRA parameters at 512^2 to condition restoration on region control R and guidance $P = (t, v)$ (Sec. 3.2.1), using the region-aware objective (Eq. (4)). Each batch uses a 4:1 mix of failure-aware and standard restoration samples. The failure-aware branch samples a localization phrase, bounding box, outline, or mask for R, with prompt-only dropout, and samples v from {empty, retrieved reference, Canny contour}. Standard restoration samples retain the stage 1 prior (Appendix C).

3.2.3 Region-aware reconstruction objective

Uniform reconstruction weighting limits flexibility in target regions, while weak supervision outside them can reduce fidelity. We weight the target less than its surroundings by converting R to a binary mask, Gaussian-smoothing it into $m : \Omega \to [0, 1]$, and reweighting only the image-space L_2 term:

$$\begin{aligned} w(\mathbf{u}) &= 1 + \alpha_{\text{in}} m(\mathbf{u}) + \alpha_{\text{out}}\big(1 - m(\mathbf{u})\big), \\ \mathcal{L}_{L_2}^{\text{region}}(\theta) &= \mathbb{E}\left[\operatorname*{mean}_{\mathbf{u}\in\Omega} w(\mathbf{u})\, \|\hat{\mathbf{x}}(\mathbf{u}) - \mathbf{x}_{\text{GT}}(\mathbf{u})\|_2^2\right]. \end{aligned} \tag{4}$$

The target region retains its standard reconstruction weight, while surrounding content receives a higher weight; Gaussian smoothing ensures a gradual transition. LPIPS and the adversarial loss remain image-wide, and $w = 1$ everywhere without localization. See Appendix C.2 for settings.

4 Experiments

Test sets. We evaluate on three test sets: (i) 120 synthetic triplets $(\mathbf{x}_{\text{LQ}}, \mathbf{x}_{\text{GT}}, v)$ constructed with the pipeline in Sec. 3.1, 30 per tier, using second-order Real-ESRGAN degradation (Wang et al., 2021) and v from semantically related samples; (ii) the paired WR-SR (Jiang et al., 2021) benchmark with



Figure 7: **Qualitative comparison on the proposed test set.** The two condition columns are the visual guidance v and the region control R of Sec. 3. FAIR follows the reference inside the target region (red box) while preserving the surroundings.

LQ–GT images and curated references; and (iii) an in-the-wild set without GT (Yin et al., 2026). Conditions are obtained automatically (Appendix D.2).

Baselines. We compare five methods that do not use v: SeeSR (Wu et al., 2024b), SUPIR (Yu et al., 2024), HYPIR (Lin et al., 2025), InstructRestore (Liu et al., 2025) (given the same region), and CODSR (Chen et al., 2025a). We further compare three reference-based methods given the same v: ReFIR+SeeSR, ReFIR+SUPIR (Guo et al., 2024), and AdaRefSR (Wang et al., 2026).

Metrics. We report 13 metrics (Table 1): PSNR, SSIM (Wang et al., 2004), FID (Heusel et al., 2017), DISTS (Ding et al., 2020), MUSIQ (Ke et al., 2021), CLIP-IQA (Wang et al., 2023), Q-Align (Wu et al., 2023), MANIQA (Yang et al., 2022), TOPIQ (Chen et al., 2024), DINO- and CLIP-Sim-Ref against v, and $\text{LPIPS}_{\text{out}}$/$\text{LPIPS}_{\text{in}}$ (Zhang et al., 2018a) outside/inside the GT mask.

4.1 Results

Among reference-based methods using the same v, FAIR reduces FID from 81.93 to 63.28 and LPIPS_{in} from 0.0368 to 0.0319 (Table 1), indicating better perceptual quality and target-region fidelity. FAIR also achieves the highest DINO-Sim-Ref and CLIP-Sim-Ref scores, showing stronger alignment with v alongside lower GT-based LPIPS_{in}. It ranks first on Q-Align and TOPIQ and achieves the lowest LPIPS_{in} in every failure tier, with the largest margin on OCR/ID (Appendix D.3); WR-SR, real-world and runtime results are in Appendix D. Fig. 7 illustrates these differences: competing methods either leave the region blurred or generate structures inconsistent with the reference, most visibly on the seal, where several baselines invent an unrelated emblem. A HYPIR-to-FLUX.2-klein-9B cascade that restores first and corrects afterward also underperforms FAIR on every reported metric (Appendix D.7).

Table 1: **Quantitative comparison on the proposed benchmark.** The best and second-best results are highlighted per metric. Reference-similarity metrics (DINO-Sim-Ref and CLIP-Sim-Ref) are reported only for methods that take reference images as input.

Metric	*Non-reference*					*Reference-based*			*Ours*
	SeeSR	SUPIR	HYPIR	Instruct Restore	CODSR	ReFIR + SeeSR	ReFIR + SUPIR	AdaRefSR	FAIR
PSNR↑	20.87	20.10	20.62	20.93	20.85	20.89	19.97	21.23	20.91
SSIM↑	0.548	0.513	0.556	0.531	0.540	0.545	0.505	0.561	0.558
$\text{LPIPS}_{\text{out}}$↓	0.300	0.340	0.272	0.356	0.278	0.301	0.389	0.277	0.243
FID↓	85.48	106.15	84.12	106.42	84.45	85.99	113.87	81.93	63.28
DISTS↓	0.1676	0.1732	0.1511	0.2070	0.1697	0.1676	0.1979	0.1621	0.1387
LPIPS_{in}↓	0.0406	0.0451	0.0367	0.0466	0.0397	0.0390	0.0489	0.0368	0.0319
MUSIQ↑	72.09	70.76	70.34	68.47	72.72	72.20	68.30	69.77	71.47
CLIPIQA↑	0.693	0.604	0.644	0.636	0.694	0.687	0.577	0.617	0.673
Q-Align↑	4.382	4.467	4.340	3.903	4.483	4.405	4.168	4.274	4.514
MANIQA↑	0.5110	0.4306	0.4412	0.4983	0.4434	0.5123	0.4211	0.4032	0.4356
TOPIQ↑	0.3978	0.3682	0.4492	0.3732	0.4177	0.3943	0.3448	0.4219	0.4655
DINO-Sim-Ref↑	–	–	–	–	–	0.412	0.386	0.448	0.468
CLIP-Sim-Ref↑	–	–	–	–	–	0.577	0.563	0.589	0.598

Table 2: **Ablation of the main design choices, using the benchmark's region annotations.**

Variant	LPIPS↓	FID↓	MUSIQ↑	CLIPIQA↑
(a) Stage-1 non-ref restoration only	0.326	80.38	65.02	0.578
(b) End-to-end (no curriculum)	0.274	64.94	70.72	0.636
(c) Frozen stage-1 LQ LoRA + new Ref LoRA	0.285	68.57	70.45	0.626
(d) Joint LQ + ref finetuning Image LoRA (LQ & Ref share)	0.274	64.63	71.44	0.658
(e) Full FAIR: (d) + reweighting	0.270	63.22	71.50	0.674

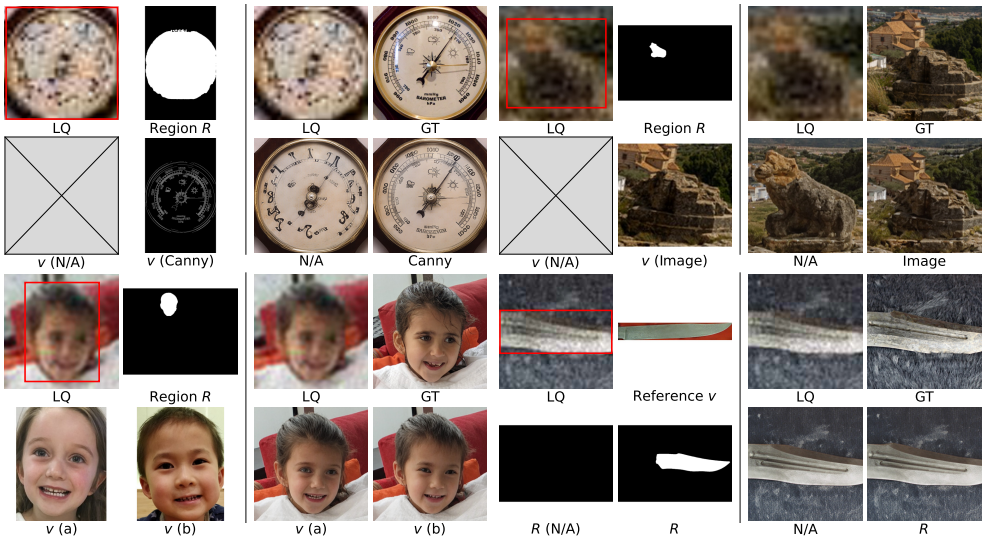

Figure 8: **Qualitative ablation of the two conditioning axes.** Restoration guidance P supplies information for recovering target content, while region control R specifies where the guidance applies.

4.2 Ablation Study

Table 2 ablates the training design of FAIR: the curriculum, LoRA sharing, joint conditioning, and region-aware reweighting; Fig. 8 and Appendix D.5 ablate the conditions themselves. Variant (a), stage-1 non-reference restoration alone, has the weakest scores across all four metrics. Replacing the two-stage curriculum with end-to-end training in variant (b) substantially improves performance, but is outperformed overall by the two-stage shared-LoRA variants (d) and (e). Variant (c) freezes the stage-1 LQ LoRA and trains a separate reference LoRA; it performs worse than variant (d), which jointly optimizes LQ and reference conditioning in a shared LoRA space. Adding the region-aware reweighting objective in Eq. (4) gives the full model (e), which is best on all four metrics.

Figure 8 shows how P and R affect restoration. A Canny contour restores the ring of graduations on the barometer dial, and a reference image restores the masonry of the ruin. Swapping one reference for another changes the recovered appearance: from the same input, a fair-haired reference yields

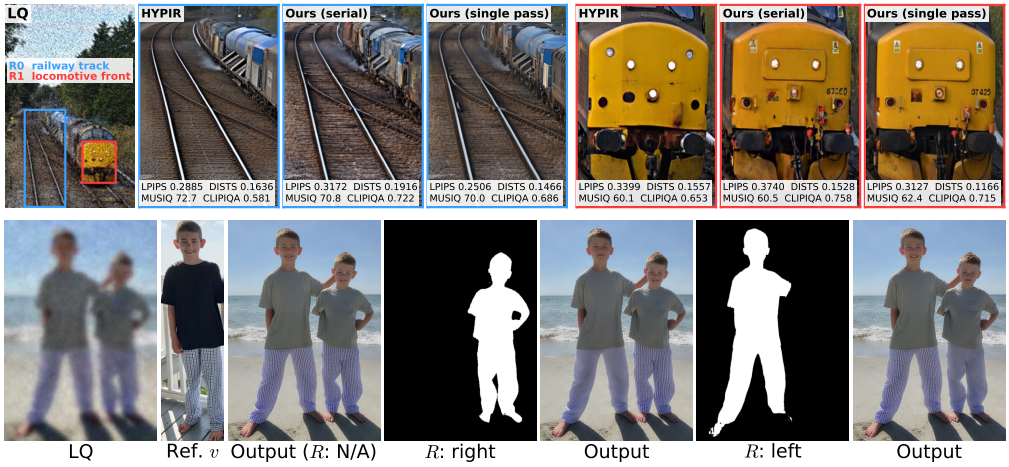

Figure 9: **Multi-region and strict region control.** Top: serial and joint restoration of two regions with separate references. Bottom: under strict region control, only the boy selected by R follows v.

lighter hair and grey eyes, a dark-haired one darker hair and browner eyes. These changes show that P guides the recovered appearance. Without R, the guidance affects content beyond the target and smooths the blade's lower edge into one curve; with R, the stepped profile survives, showing that the guidance location is controlled separately.

4.3 Multiple Regions with Per-Region References

For an image containing multiple target regions that require different references, FAIR can restore them jointly. Specifically, we combine the region masks and tile their corresponding references. Compared with processing one region per pass, which repeatedly re-encodes the previous restoration, joint processing avoids the accumulation of artifacts across passes (Almog et al., 2025). Figure 9 compares the two strategies on target-region crops. Under serial processing, restoring the second region unintentionally alters the first, whereas joint processing preserves both. The bottom row shows strict region control: without R, both boys follow the plaid in v; with R, only the selected boy does.

4.4 Blinded Subjective Study

Because IQA metrics may not fully capture human perception, twenty observers independently selected the top two among nine anonymized outputs for each of 10 cases, evaluating overall quality, target correction, and non-target preservation separately. FAIR attains 77.5–79.5% Top-2 rates compared with 40.0–45.0% for the strongest baseline, HYPIR (Table 3), and ranks first among all methods on each dimension (Appendix E).

Table 3: **User study.** Parentheses give 95% Wilson intervals; HYPIR is second to FAIR.

Dimension	Ours		HYPIR
Overall quality	**79.5%**	(73.4–84.5)	40.0%
Target correction	**77.5%**	(71.2–82.7)	44.0%
Non-target preservation	**78.5%**	(72.3–83.6)	45.0%

5 Conclusion

We address the last-mile problem in generative image restoration, where localized failures can make an otherwise plausible result unacceptable. FAIR predicts failure-prone regions from the low-quality input, prepares restoration guidance $P = (t, v)$ for their needs, and applies it through region control R during restoration. A $67k$-region failure-aware dataset, localized multimodal conditioning, and a two-stage region-aware curriculum support this approach. Across synthetic, reference-based, and in-the-wild settings, FAIR improves perceptual quality, reference alignment, and local controllability. A blinded study also favors FAIR on overall quality, target correction, and non-target preservation. These results show the value of identifying where restoration needs additional evidence and preparing the corresponding guidance in advance.

ETHICS STATEMENT

This work involves human subjects and the editing of identity-bearing image content, and we address both. The blinded subjective study (Appendix E) collected only anonymous preference judgements: no personal data, demographics, or free-text responses were recorded, and observers could stop at any point. Training and evaluation images come from LSDIR, Unsplash, and FFHQ under their respective licences; we add region-level failure annotations and release no new imagery of identifiable individuals beyond what those datasets already contain.

Our method edits faces, text, and logos, which carries a risk of producing plausible but incorrect detail, or of steering an identity toward a retrieved exemplar. Restored outputs are therefore reconstructions, not evidence: they should be disclosed as AI-assisted and should not be relied upon in forensic, medical, journalistic, or legal settings.

REPRODUCIBILITY STATEMENT

Sec. 3.1 and Appendix B give the full data construction pipeline, including the annotation prompts, the SAM3 mask filtering, and the reference retrieval and filtering steps. Appendix C lists the architecture, LoRA placement, loss weights, optimizer settings, batch sizes, and hardware for both training stages; Sec. 3.2.3 defines the region-aware objective and its parameters. The evaluation protocol, metric definitions, and the blinded-study design are specified in Sec. 4 and Appendix E. We will release the annotation pipeline, the failure-region dataset, and inference code.

AI USE STATEMENT

We used generative AI tools as writing aids: for language polishing of author-written text and for assistance with LaTeX formatting. We also used them as coding assistants for parts of the experimental and evaluation scripts. We did not use generative AI to generate research ideas, to design the method, to produce or alter any experimental result, or to write the scientific claims of this paper. All AI-assisted text and code were reviewed by the authors, and all reported numbers were produced by the pipelines described above and verified against the raw logs. We take responsibility for the final content of this work.

REFERENCES

Gal Almog, Ariel Shamir, and Ohad Fried. Reed-vae: Re-encode decode training for iterative image editing with diffusion models. *Computer Graphics Forum*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Jiezhang Cao, Jingyun Liang, Kai Zhang, Yawei Li, Yulun Zhang, Wenguan Wang, and Luc Van Gool. Reference-based image super-resolution with deformable attention transformer. In *European conference on computer vision*, pp. 325–342. Springer, 2022.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. TOPIQ: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024. doi: 10.1109/TIP.2024.3378466.

Hao Chen, Junyang Chen, Jinshan Pan, and Jiangxin Dong. Bridging fidelity-reality with controllable one-step diffusion for image super-resolution. *arXiv preprint arXiv:2512.14061*, 2025a.

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, 2025b.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22367–22377, 2023.

Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11065–11074, 2019.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

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

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

Hang Guo, Tao Dai, Zhihao Ouyang, Taolin Zhang, Yaohua Zha, Bin Chen, and Shu-tao Xia. Refir: Grounding large restoration models with retrieval augmentation. *Advances in Neural Information Processing Systems*, 37:46593–46621, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yuming Jiang, Kelvin CK Chan, Xintao Wang, Chen Change Loy, and Ziwei Liu. Robust reference-based super-resolution via c2-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2103–2112, 2021.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. `https://bfl.ai/blog/flux-2`, 2025.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

Yawei Li, Kai Zhang, Jingyun Liang, Jiezhang Cao, Ce Liu, Rui Gong, Yulun Zhang, Hao Tang, Yun Liu, Denis Demandolx, et al. Lsdir: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1775–1787, 2023.

Jie Liang, Hui Zeng, and Lei Zhang. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5657–5666, 2022.

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.

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

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

Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration. *ACM Transactions on Graphics (TOG)*, 44(6):1–21, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Shuaizheng Liu, Jianqi Ma, Lingchen Sun, Xiangtao Kong, and Lei Zhang. Instructrestore: Region-customized image restoration with human instructions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Liying Lu, Wenbo Li, Xin Tao, Jiangbo Lu, and Jiaya Jia. Masa-sr: Matching acceleration and spatial adaptation for reference-based image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6368–6377, 2021.

Yiqun Mei, Yuchen Fan, and Yuqian Zhou. Image super-resolution with non-local sparse attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3517–3526, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Haoze Sun, Wenbo Li, Jianzhuang Liu, Haoyu Chen, Renjing Pei, Xueyi Zou, Youliang Yan, and Yujiu Yang. Coser: Bridging image and language for cognitive super-resolution. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 25868–25878, 2024.

Unsplash. The Unsplash Dataset. `https://github.com/unsplash/datasets`, 2020.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1905–1914, 2021.

Yuan Wang, Yuhao Wan, Siming Zheng, Bo Li, Qibin Hou, and Peng-Tao Jiang. Trust but verify: Adaptive conditioning for reference-based diffusion super-resolution via implicit reference correlation modeling. *arXiv preprint arXiv:2602.01864*, 2026.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17683–17693, 2022.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Gang Wu, Junjun Jiang, Kui Jiang, Xianming Liu, and Liqiang Nie. Dswinir: Rethinking window-based attention for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

Rongyuan Wu, Lingchen Sun, Zhiyuan Ma, and Lei Zhang. One-step effective diffusion network for real-world image super-resolution. *Advances in Neural Information Processing Systems*, 37: 92529–92553, 2024a.

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, 2024b.

Jiaqi Yan, Shuning Xu, Xiangyu Chen, Dell Zhang, Jie Tang, Gangshan Wu, and Jie Liu. Rasr: Retrieval-augmented super resolution for practical reference-based image restoration. *arXiv preprint arXiv:2508.09449*, 2025.

Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5791–5800, 2020.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022.

Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In *European conference on computer vision*, pp. 74–91. Springer, 2024.

Xiang Yin, Jinfan Hu, Zhiyuan You, Kainan Yan, Yu Tang, Chao Dong, and Jinjin Gu. How far have we gone in generative image restoration? a study on its capability, limitations and evaluation practices. *arXiv preprint arXiv:2603.05010*, 2026.

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, 2024.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018a.

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

Zhifei Zhang, Zhaowen Wang, Zhe Lin, and Hairong Qi. Image super-resolution by neural texture transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7982–7991, 2019.

APPENDIX OVERVIEW

The appendix is organized as follows. Appendix A compares FAIR with semantic-guided and reference-based restoration methods. Appendix B details the three-stage failure-aware data pipeline and provides additional dataset examples and statistics. Appendix C reports the architecture, optimization, and training configuration. Appendix D presents additional quantitative and qualitative results, including a comparison with a cascade of off-the-shelf models. Appendix E describes the blinded subjective study, and Appendix F evaluates automatic condition acquisition, component scale, reference mismatch, and run-to-run variation.

A METHOD COMPARISON

We compare FAIR with representative diffusion-based restoration methods along three axes: *auxiliary input* (what external information is supplied), *region control* (whether and how that information is spatially localized), and *inference efficiency* (steps per image). Table 4 summarizes the comparison.

Semantic-guided methods (Yang et al., 2024; Wu et al., 2024b; Yu et al., 2024; Sun et al., 2024; Liu et al., 2025; Lin et al., 2025; Chen et al., 2025a) supply auxiliary inputs ranging from text prompts and captions to segmentation masks. Some provide spatial control, but they do not select restoration guidance by predicting a local failure type. Reference-based methods (Guo et al., 2024; Yan et al., 2025; Wang et al., 2026) provide external visual evidence, typically without binding it to a predicted failure region.

FAIR instead predicts failure-prone regions and their types from the low-quality input, prepares the corresponding restoration guidance, and applies it through explicit region control.

Table 4: **Comparison of diffusion-based restoration methods.** Auxiliary input, region control, and inference steps per image.

Method	Auxiliary Input	Region Control & Alignment	Steps
Semantic-guided diffusion restoration			
PASD (Yang et al., 2024)	Detector + caption	Implicit (cross-attn)	20
SeeSR (Wu et al., 2024b)	Text prompt	Implicit (cross-attn)	50
SUPIR (Yu et al., 2024)	Text prompt	Implicit (cross-attn)	50
CoSeR (Sun et al., 2024)	Cognitive embedding	Implicit (cross-attn)	50
InstructRestore (Liu et al., 2025)	Instruction	Instruction-level (implicit)	20
HYPIR (Lin et al., 2025)	Text prompt	Implicit (cross-attn)	1
CODSR (Chen et al., 2025a)	Tag + segmentation mask	Region-adaptive prior + mask	1
Reference-based diffusion restoration			
ReFIR (Guo et al., 2024)	Retrieved reference	Global (full-image cross-attn)	50
RASR (Yan et al., 2025)	Category-retrieved reference	Global (full-image attn)	50
AdaRefSR (Wang et al., 2026)	Retrieved reference	Adaptive cond. (implicit)	1
FAIR	**Multi-modal (text / reference / contour)**	**Explicit region control (self-attn) (phrase / bbox / outline / mask)**	**1**

B DATA CONSTRUCTION DETAILS

We provide additional details for each stage of the failure-aware data construction pipeline described in Sec. 3.1. The guidance types of Fig. 2 map onto $P = (t, v)$ as in Table 5: semantic instructions form t; reference images, geometry prompts (Canny contours) and identity correspondences each enter as v; parametric controls, such as the noise and sharpness levels, are inference settings rather than part of P.

B.1 STAGE 1: DATA GENERATION AND SEED ANNOTATION

Source images. We collect ground-truth images from two sources: LSDIR (Li et al., 2023), providing 84,991 high-resolution natural images curated from Flickr (we use the official training split), and the Unsplash Lite dataset (Unsplash, 2020), a curated collection of $\sim$25k photographer-uploaded high-quality images. All images are resized so that the shorter edge has 1024 pixels, then center-cropped to 1024×1024 patches.

Table 5: **Summary of conditions and notation.** At the task level, R specifies where additional guidance is applied and P specifies what guidance is used. At the model input, P is represented by text tokens t or visual tokens v.

Symbol	Name	Source	Representation at input	Primarily serves
$\mathbf{x}_{\mathrm{LQ}}$	degraded input	degradation of $\mathbf{x}_{\mathrm{GT}}$	VAE latent, image branch	all tiers
$\mathbf{x}_{\mathrm{GT}}$	clean target	LSDIR, Unsplash, or FFHQ image	supervision only	all tiers
R	region control	localization phrase, box, outline, or mask; $\varnothing$ for prompt-only dropout	spatial condition	all tiers
P	restoration guidance	text instruction, retrieved reference, or contour	t or v	selected by F
t	textual instruction	text instruction P	text-stream tokens	text guidance
v	visual guidance	reference or contour P	VAE tokens, image branch	visual guidance
$F = (F_{\text{tier}}, F_{\text{sub}})$	failure label	fine-tuned VLM	selects P	—
$\tilde{\mathbf{x}}$	coarse restoration	existing restoration model	**offline only**	—
$\hat{\mathbf{x}}$	restored output	$f_\theta(\mathbf{x}_{\mathrm{LQ}}, R, P)$	model output	all tiers

Degradation pipeline. LQ inputs are synthesized using the second-order degradation pipeline of Real-ESRGAN (Wang et al., 2021), which sequentially applies two rounds of (blur, downsampling, noise, JPEG compression) with parameters sampled from the official Real-ESRGAN configuration. This pipeline is widely adopted by recent restoration methods (Wu et al., 2024b;a; Lin et al., 2025) and produces LQ images that approximate real-world degradations.

Coarse restorations. Each $\mathbf{x}_{\mathrm{LQ}}$ is processed by multiple state-of-the-art restoration models (Lin et al., 2024; Yu et al., 2024; Wu et al., 2024b; Lin et al., 2025). Their outputs $\tilde{\mathbf{x}}$ expose realistic local failure modes rather than synthetic perturbations.

Seed annotation prompts. We use Qwen3-VL-8B (Bai et al., 2025) with the following prompt structure to identify failure regions from $(\mathbf{x}_{\mathrm{LQ}}, \tilde{\mathbf{x}}, \mathbf{x}_{\mathrm{GT}})$ triplets.

First-pass prompt (failure localization and typing).

```
You are given three images: a low-quality input (LQ),
its restored result (Restoration), and the corresponding
ground-truth image (GT). Compare the restored result
with both the LQ input and GT, and identify semantically
meaningful regions where restoration fails.
For each detected failure, return a JSON object with the
following fields:
  • region_description: a brief natural-language
    description of the region (e.g., "the woman's face on
    the left", "the text on the street sign").
  • failure_type: one of {OCR/ID, Object Structure,
    Pattern/Texture, Subjective Feeling}.
  • severity: one of {minor, moderate, severe}.
  • global_label: good if the restored result contains
    no notable local failures relative to the GT; gt_skip
    if the GT itself is poor quality and cannot be used as
    reference; otherwise ok.
If no failures are detected, return an empty list.
```

Second-pass prompt (textual re-classification). The second pass processes each detected failure independently, using only its textual description (without re-accessing the images) to re-classify the failure type. This pass corrects inconsistencies in the initial typing, particularly for borderline cases where Qwen3-VL produced overly broad or mis-categorized labels in the first pass.

```
Given the following failure region description, re-classify
it into one of the following failure types: {OCR/ID, Object
Structure, Pattern/Texture, Subjective Feeling}.
Description: {region_description}
```

```
Provide your reasoning step by step before giving the
final classification.  Return JSON: {"reasoning":  "...",
"failure_type":  "..."}
```

Filtering. We keep only samples with `global_label = ok` and at least one detected failure; `original_failure_type` retains the first-pass label for ablation.

Manual screening and fine-tuning. We manually accept or reject the seed predictions and retain 1K correct region–type annotations. The verified $(\mathbf{x}_{\mathrm{LQ}}, \tilde{\mathbf{x}}, \mathbf{x}_{\mathrm{GT}})$ triplets are then used to fine-tune Qwen3-VL-8B with rank-64 LoRA.

B.2 Stage 2: Failure Annotation and Region Construction

Large-scale offline annotation. The fine-tuned annotator processes the full pool using the same input triplet and predicts a localization phrase, failure label, and validity flag for each failure. Ground truth and coarse restorations are used only during offline data construction; the automatic inference mode receives $\mathbf{x}_{\mathrm{LQ}}$ alone.

Description-to-phrase translation. Region descriptions from Stage 1 are often long, contain Chinese fragments, or include qualifiers (e.g., "the slightly blurred face of the woman on the right"). For SAM3 to produce a clean mask, we first translate each description into a concise English noun phrase using Qwen3-VL:

```
Convert the following region description into a concise
English noun phrase suitable for open-vocabulary
segmentation.  The phrase should be 1-6 words and refer
unambiguously to the object or area.
Description:  {region_description}
Output only the noun phrase, with no additional text.
```

SAM3 segmentation. The translated noun phrase, together with the HQ image, is fed into SAM3 (Carion et al., 2025) (text-prompted mode) to obtain a binary mask. We use SAM3's default confidence threshold and retain the mask with the highest confidence score for each region.

Region filtering and control derivation. We filter out regions whose SAM3 mask area falls outside $[\rho_{\text{min}}, \rho_{\text{max}}]$ relative to the full image area, using $\rho_{\text{min}} = 0.5\%$ to remove likely segmentation noise and $\rho_{\text{max}} = 50\%$ to exclude non-local regions. A bounding box and closed outline are derived from each retained mask. With the localization phrase, these instantiate the four region-control formats R.

B.3 Stage 3: Reference Retrieval and Guidance Construction

Global reference gallery. We partition the LSDIR and Unsplash training images into overlapping multi-scale tiles at relative scales $\{0.33, 0.50, 0.67\}$ with stride ratio 0.5. DINOv3 (Siméoni et al., 2025) features extracted from these tiles form a global FAISS index containing 4.13M vectors merged from 18 shards.

Region-aware retrieval and ranking. For each annotated region, its local GT crop queries the global index. We retrieve 1,024 DINO nearest neighbors, exclude candidates from the same source image, and rerank the top 120 using

$$s = 0.45 s_{\mathrm{sem}} + 0.55 s_{\mathrm{vis}}, \tag{5}$$

where s_{vis} is the DINO cosine similarity to the annotated local region and s_{sem} is the CLIP (Radford et al., 2021) image–text similarity between the candidate and region description. We retain candidates with $s \geq 0.70$ and $s_{\mathrm{vis}} \in [0.50, 0.98)$; the upper bound removes near-duplicate photographs.

Semantic recropping and filtering. Because grid tiles may truncate the target or contain mostly background, we prompt SAM3 (Carion et al., 2025) with the region description on the candidate source image and recrop the predicted instance box, subject to a minimum area ratio of 1%. SAM3

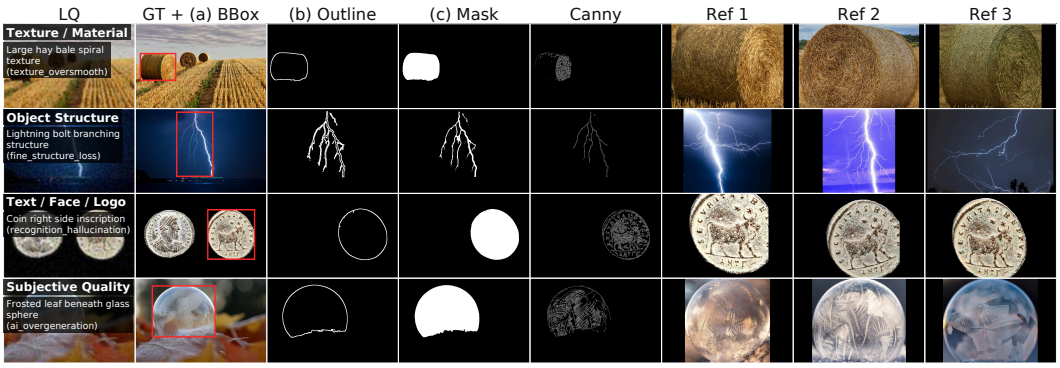

Figure 10: **Additional samples from our failure-aware dataset, one per tier.** Each row shows the LQ–GT pair, the four region-control formats, the matched $P = (t, v)$, and the failure label F.

serves only as a recropper: if no instance is found, we retain the original tile, since its no-instance decision is less reliable (68.4% accuracy) than unconditional retention (77.9%). Qwen3-VL-8B then filters candidates for semantic agreement with the region description, achieving 86.7% validation accuracy. At most one reference is retained from each source image, yielding three references from distinct images whenever possible.

Reference-slot routing. Across failure tiers, the visual slot of P is routed among a reference image, a Canny edge map, and black-image dropout with probabilities 0.760, 0.068, and 0.172, respectively.

B.4 Dataset Statistics

Table 6 summarizes the final dataset composition.

Table 6: **Final failure-aware training dataset statistics.** Region counts are reported after Stage 3 filtering.

Statistic	Count
Stage 1 candidate regions	169,257
Final training regions (post-filter)	66,989
Failure type distribution	
Pattern / Texture	63.9%
Object Structure	21.0%
OCR / ID	10.1%
Subjective Feeling	5.0%

C Training and Implementation Details

C.1 Stage 1: Restoration Pre-training

Stage 1 establishes a general LQ-to-HQ prior without reference conditioning. We adapt FLUX.2-klein-9B with LoRA (rank 64) on all linear projections per transformer block (text stream frozen), and train at 1024×1024 on a three-source mix of Unsplash regular (66%), Unsplash sharp (30%), and FFHQ (4%), degraded with the second-order Real-ESRGAN pipeline (Wang et al., 2021). Captions are split $0.46/0.54$ between the null prompt `"Restore"` and per-image LLaVA-style dense captions. Optimization uses RMSprop ($\eta_G = \eta_D = 10^{-6}$) with batch size 64 across 32 H200 GPUs (per-GPU batch 2) for 16,000 iterations. Training uses fixed-timestep single-step reconstruction at $t = 500$ with noise augmentation $\sigma_{\text{aug}} \sim \mathcal{U}[200, 800]$: the predicted velocity is used only to form the one-step clean latent $z_0 = z_t - \sigma_t v_\theta$, which is decoded to $\hat{\mathbf{x}}$. No target velocity is constructed and no separate velocity-supervision or diffusion loss is applied. The generator loss of Eq. (3) therefore acts entirely on the decoded image and combines image-space L_2, LPIPS (Zhang et al., 2018a), and a GAN loss using a ConvNeXt discriminator. We set $\lambda_{L_2} = \lambda_{\text{LPIPS}} = 10$, $\lambda_{\text{GAN}} = 0.25$. The discriminator uses R1 regularization ($\gamma = 100$) enabled after 1,000 iterations.

Table 7: **Quantitative comparison on the real-world test set.**

Metric	*Non-reference*						*Reference-based*			*Ours*
	SeeSR	SUPIR	HYPIR	Instruct Restore	CODSR	Nano Banana	ReFIR + SeeSR	ReFIR + SUPIR	AdaRefSR	FAIR
MUSIQ↑	66.01	56.22	64.72	65.06	69.93	45.99	66.07	55.70	59.65	69.62
CLIPIQA↑	0.633	0.433	0.601	0.686	0.653	0.402	0.628	0.428	0.514	0.671
Q-Align↑	3.803	3.535	3.883	3.706	4.171	3.339	3.815	3.445	3.727	4.068
DINO-Sim-Ref↑	–	–	–	–	–	–	0.363	0.364	0.363	0.369
CLIP-Sim-Ref↑	–	–	–	–	–	–	0.531	0.533	0.534	0.542

C.2 Stage 2: Region-Guided Fine-Tuning

We fine-tune for 30,000 iterations at 512×512 in `bf16` with the same LoRA target set. Each batch mixes failure-aware data (our 66,989-region dataset with full reference conditioning) and Real-ESRGAN-degraded restoration data (Unsplash and FFHQ) in the 4:1 ratio of Sec. 3. The conditioning format is sampled as $\{\texttt{prompt_only}, \texttt{prompt_loc}, \texttt{bbox}, \texttt{outline}, \texttt{mask}\}$ with probabilities $\{0.10, 0.15, 0.25, 0.25, 0.25\}$. The `prompt_only` mode sets $R = \varnothing$, while the localization phrase in `prompt_loc` enters t and leaves the binary region map blank. Bounding boxes and outlines are filled to form binary region maps; masks are used directly. The map and v are VAE-encoded and appended to the noisy-target and $\mathbf{x}_{\mathrm{LQ}}$ tokens as separate reference images, each with its own positional offset, so that every condition reaches the target through the backbone's joint self-attention over a single token stream. The visual guidance v follows the reference-slot routing in Appendix B.3. The training loss follows Stage 1 with $\lambda_{L_2} = \lambda_{\mathrm{LPIPS}} = 10$ and $\lambda_{\mathrm{GAN}} = 0.25$, plus the region-aware reweighting

$w = 1 + 4(1 - m)$ of Eq. (4). The reweighting is applied to the image-space L_2 term *only*; LPIPS and the adversarial loss remain image-wide, since spatially masking them would weaken the perceptual and realism constraints in the target region. As in stage 1, all losses act on the decoded image. Optimization keeps RMSprop ($\eta_G = \eta_D = 10^{-6}$) as in stage 1.

D Additional Experimental Results

D.1 Real-World and WR-SR Benchmarks

The synthetic-set findings of Sec. 4.1 carry over to both real-world settings. In Tables 7–10 and 12, the best and second-best results are shown in red and blue. On the in-the-wild real-world test set of 207 images (Yin et al., 2026), shown in Table 7, FAIR is best on both reference-similarity metrics and second on MUSIQ, CLIPIQA, and Q-Align. CODSR and InstructRestore score higher on some no-reference metrics. Nano Banana, a general-purpose commercial image model, scores lower on the reported metrics. On the 80 WR-SR test images (Jiang et al., 2021), shown in Table 8, FAIR is best on five of ten metrics, SSIM, LPIPS, FID and both reference-side similarities, the latter by wide margins (+0.056 DINO, +0.030 CLIP). PSNR is within 0.24 dB of AdaRefSR, while CODSR and the SeeSR-based methods lead the no-reference scores. Runtime, reported in the last row of Table 8, is 3.16 s, competitive with AdaRefSR (2.30 s). Figure 13 shows qualitative results on both sets.

D.2 Benchmark Composition

The synthetic benchmark holds 120 triplets, 30 per failure tier, built by the Sec. 3.1 pipeline; none of its regions is among the 66,989 training regions. References are constructed by the same retrieval procedure as in training (Appendix B.3). $\mathrm{LPIPS}_{\mathrm{in}}$ and $\mathrm{LPIPS}_{\mathrm{out}}$ are computed on the full image after replacing the pixels outside or inside the GT failure mask, respectively, with GT.

Conditions at test time. Apart from the references of WR-SR, which ships them as a public reference-based super-resolution benchmark, all conditions in the reported results are produced by the automatic pipeline; (R, t, v) can equally be supplied by a user. On the synthetic benchmark, R is predicted from $\mathbf{x}_{\mathrm{LQ}}$ by the pipeline of Appendix F, while t and v are produced by the automatic annotation and retrieval of Sec. 3.1. On WR-SR, v is the reference paired with each image by the

Table 8: **Quantitative comparison on the WR-SR test set.**

Metric	*Non-reference*					*Reference-based*			*Ours*
	SeeSR	SUPIR	HYPIR	Instruct Restore	CODSR	ReFIR + SeeSR	ReFIR + SUPIR	AdaRefSR	FAIR
PSNR↑	21.72	20.77	21.10	21.24	21.42	21.85	21.04	22.04	21.80
SSIM↑	0.570	0.528	0.565	0.558	0.547	0.569	0.539	0.581	0.591
LPIPS↓	0.349	0.394	0.335	0.430	0.333	0.341	0.394	0.298	0.262
FID↓	64.64	85.50	76.93	121.71	62.20	61.22	77.55	48.85	38.90
MUSIQ↑	72.09	69.94	69.66	69.53	72.92	72.42	70.06	70.21	71.37
CLIPIQA↑	0.711	0.624	0.664	0.606	0.724	0.715	0.623	0.669	0.703
Q-Align↑	4.40	4.30	4.19	3.79	4.58	4.44	4.30	4.36	4.48
DINO-Sim-Ref↑	–	–	–	–	–	0.689	0.663	0.768	0.824
CLIP-Sim-Ref↑	–	–	–	–	–	0.750	0.727	0.782	0.811
Runtime/s↓	34.91	37.61	5.76	17.58	1.71	35.06	40.68	2.30	3.16

benchmark, R covers the whole image, and t is a Qwen3-VL caption of $\mathbf{x}_{\mathrm{LQ}}$. On the real-world set, which has no GT, R and t are predicted from $\mathbf{x}_{\mathrm{LQ}}$, and v is a reference retrieved with $\mathbf{x}_{\mathrm{LQ}}$ from the gallery of Appendix B.3.

D.3 Results by Failure Tier

The test set holds 30 regions per tier, so the aggregate numbers of Table 1 are not dominated by the tier that is most frequent in training (Pattern/Texture, 63.9% of regions). Table 9 breaks $\mathrm{LPIPS}_{\mathrm{in}}$ down by tier. FAIR leads every tier, and by the widest margin on OCR/ID, the identity-critical case that motivates the taxonomy: 0.0103 against 0.0172 for the best baseline, a 40% reduction, against 7% on Pattern/Texture. Outside the target, FAIR also has the lowest $\mathrm{LPIPS}_{\mathrm{out}}$ in every tier, so the in-region gains do not come at the expense of the rest of the image.

Table 9: **$\mathrm{LPIPS}_{\mathrm{in}}$ and $\mathrm{LPIPS}_{\mathrm{out}}$ by failure tier (30 regions each).**

Tier	ReFIR+SeeSR	HYPIR	CODSR	AdaRefSR	FAIR
$LPIPS_{in}$↓					
OCR / ID	0.0190	0.0188	0.0203	0.0172	0.0103
Object Structure	0.0331	0.0303	0.0317	0.0302	0.0264
Pattern / Texture	0.0611	0.0573	0.0624	0.0586	0.0534
Subjective Feeling	0.0430	0.0403	0.0443	0.0411	0.0377
$LPIPS_{out}$↓					
OCR / ID	0.306	0.279	0.283	0.280	0.238
Object Structure	0.286	0.272	0.275	0.267	0.242
Pattern / Texture	0.302	0.265	0.263	0.278	0.247
Subjective Feeling	0.309	0.270	0.293	0.284	0.245

D.4 Equal-Information Comparison

To separate the model from the extra conditions, we run FAIR with $\mathbf{x}_{\mathrm{LQ}}$ alone: R and v are left empty and the instruction is the generic "Restore"; checkpoint and inference settings match Table 1. Against the five baselines without v (Table 10), FAIR still leads six of eleven metrics, including $\mathrm{LPIPS}_{\mathrm{out}}$, $\mathrm{LPIPS}_{\mathrm{in}}$, FID and DISTS, so its advantage does not come from the conditions alone. Adding R, t and v then lowers FID from 71.79 to 63.28 and $\mathrm{LPIPS}_{\mathrm{in}}$ from 0.0353 to 0.0319.

D.5 Removing the Visual Guidance at Test Time

Removing v while keeping R and t raises $\mathrm{LPIPS}_{\mathrm{in}}$ by 68% on OCR/ID but only 1.6% on Subjective Feeling (Table 11). The harder the tier, the more the external evidence matters, as the taxonomy of Fig. 2 assumes; Subjective Feeling, which it serves with an instruction alone, barely needs a reference.

D.6 Additional Qualitative Comparisons

Figures 11 and 12 show further cases from the proposed test set, and Figures 13 and 14 show cases from the real-world test set and WR-SR. Each case gives the LQ input, the visual guidance v, the

Table 10: **Equal-information comparison on the synthetic benchmark.** FAIR (LQ only) receives $\mathbf{x}_{\mathrm{LQ}}$ alone; best and second-best among the methods without v are highlighted. The last column repeats FAIR with its conditions (Table 1).

Metric	SeeSR	SUPIR	HYPIR	Instruct Restore	CODSR	FAIR (LQ only)	FAIR
PSNR↑	20.87	20.10	20.62	20.93	20.85	20.80	20.91
SSIM↑	0.548	0.513	0.556	0.531	0.540	0.551	0.558
$\mathrm{LPIPS}_{\mathrm{out}}$↓	0.300	0.340	0.272	0.356	0.278	0.251	0.243
FID↓	85.48	106.15	84.12	106.42	84.45	71.79	63.28
DISTS↓	0.1676	0.1732	0.1511	0.2070	0.1697	0.1495	0.1387
$\mathrm{LPIPS}_{\mathrm{in}}$↓	0.0406	0.0451	0.0367	0.0466	0.0397	0.0353	0.0319
MUSIQ↑	72.09	70.76	70.34	68.47	72.72	71.04	71.47
CLIPIQA↑	0.693	0.604	0.644	0.636	0.694	0.663	0.673
Q-Align↑	4.382	4.467	4.340	3.903	4.483	4.485	4.514
MANIQA↑	0.5110	0.4306	0.4412	0.4983	0.4434	0.4249	0.4356
TOPIQ↑	0.3978	0.3682	0.4492	0.3732	0.4177	0.4557	0.4655

Table 11: **$\mathrm{LPIPS}_{\mathrm{in}}$ with and without v by failure tier (30 regions each).**

Tier	with v	without v	p
OCR / ID	0.0103	0.0173 (+68%)	4×10^{-6}
Object Structure	0.0264	0.0291 (+10%)	0.022
Pattern / Texture	0.0534	0.0553 (+3.6%)	0.047
Subjective Feeling	0.0377	0.0383 (+1.6%)	0.22

Table 12: **FAIR compared with a two-stage cascade.**

	PSNR↑	SSIM↑	$\mathrm{LPIPS}_{\mathrm{out}}$↓	DISTS↓	$\mathrm{LPIPS}_{\mathrm{in}}$↓	MUSIQ↑
HYPIR → FLUX.2-klein-9B	17.28	0.478	0.322	0.162	0.0482	70.83
FAIR	20.93	0.559	0.242	0.139	0.0330	71.37

region control R and the textual instruction t, followed by the outputs of the compared methods. As in Figure 7, FAIR follows the reference inside the target region while preserving the surroundings.

D.7 Cascading Off-the-Shelf Tools

We compare FAIR with a cascade of HYPIR followed by unmodified FLUX.2-klein-9B on the same inputs. In Fig. 15, the first model's output has lost cues needed for faithful restoration; the second model then copies the reference plate, producing plausible but incorrect content. FAIR leads the cascade on every metric in Table 12, by 3.6 dB PSNR and 0.080 $\mathrm{LPIPS}_{\mathrm{out}}$, and on 115 and 113 images respectively ($p < 10^{-19}$, Wilcoxon signed-rank). This comparison supports supplying region and reference guidance together during restoration.

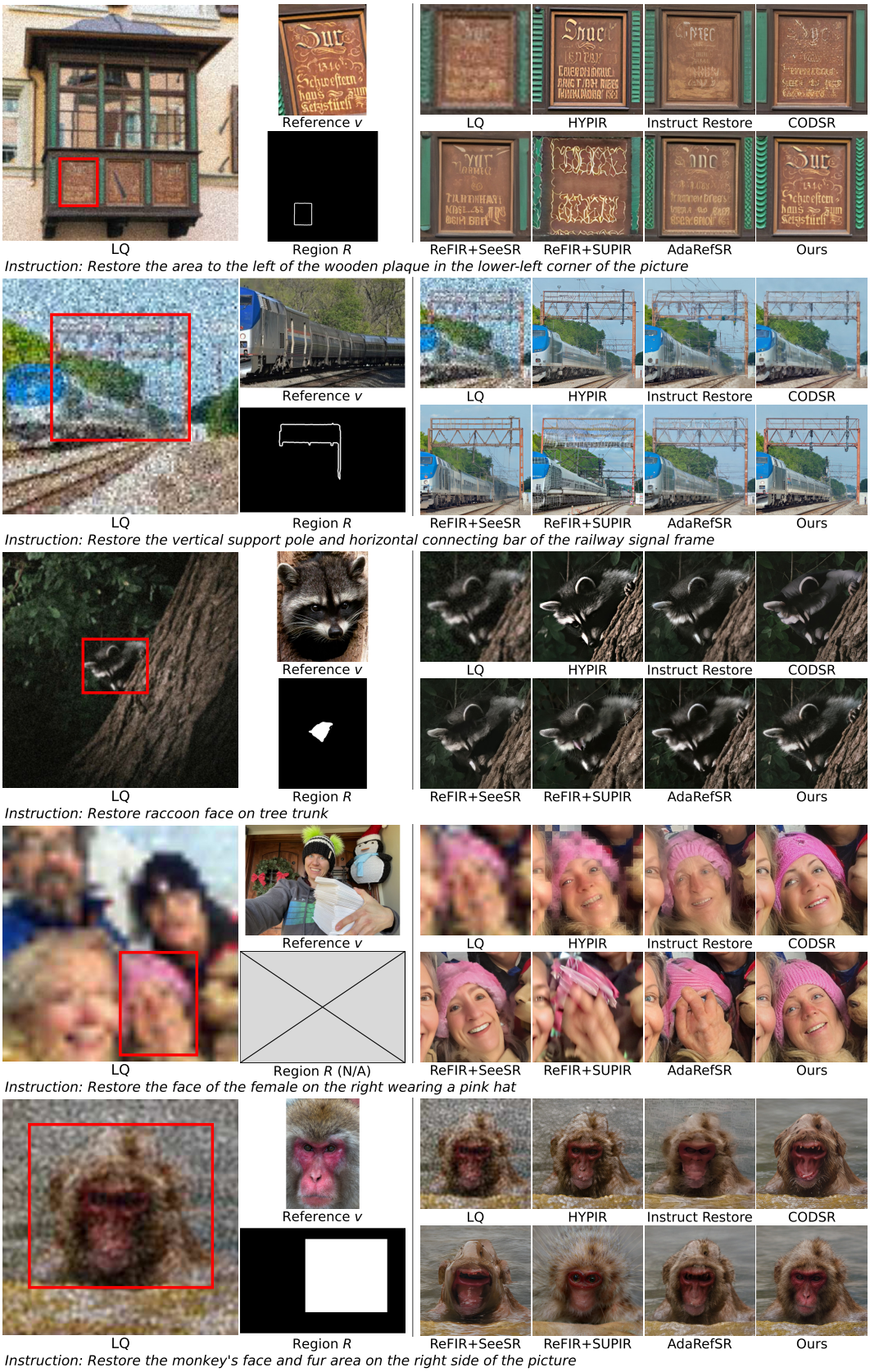

Figure 11: **Additional qualitative comparisons on the proposed test set (1/2).** Each case shows the LQ input with the target region (red box), the visual guidance v, the region control R, and the textual instruction t, followed by the outputs of the compared methods.

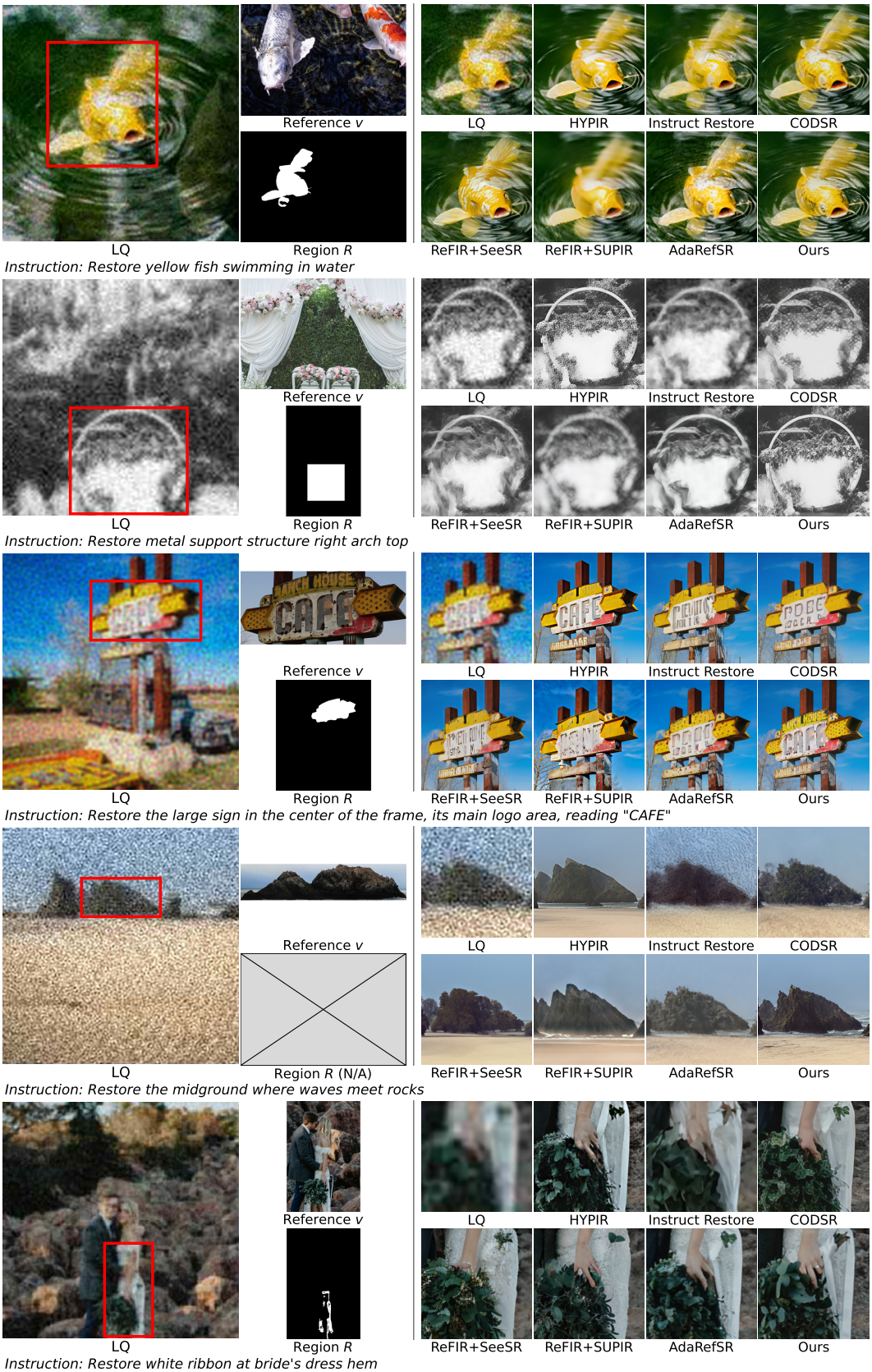

Figure 12: **Additional qualitative comparisons on the proposed test set (2/2).** Each case shows the LQ input with the target region (red box), the visual guidance v, the region control R, and the textual instruction t, followed by the outputs of the compared methods.

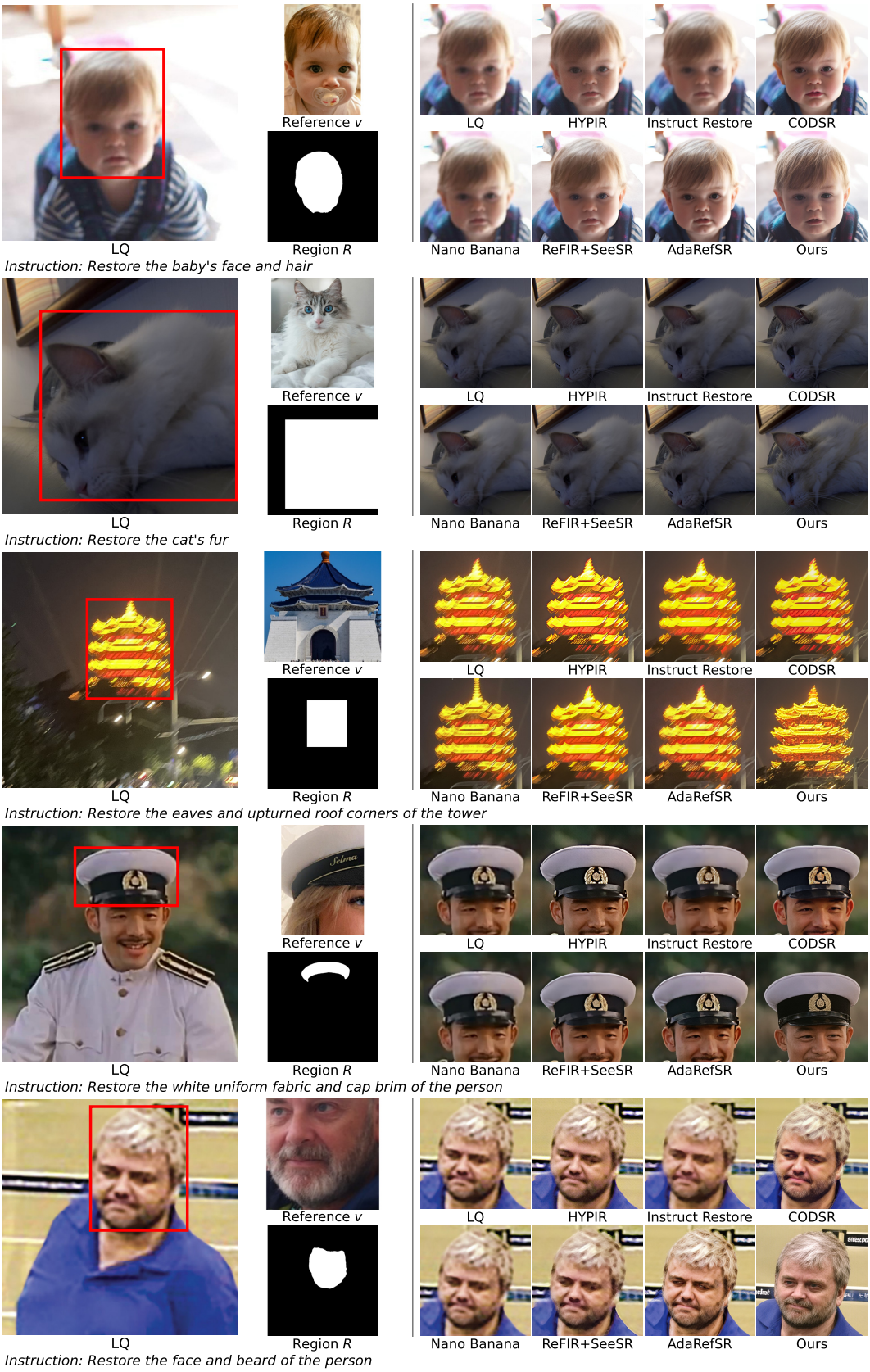

Figure 13: **Additional qualitative comparisons on the real-world test set.** Each case shows the LQ input with the target region (red box), the visual guidance v, the region control R, and the textual instruction t, followed by the outputs of the compared methods.

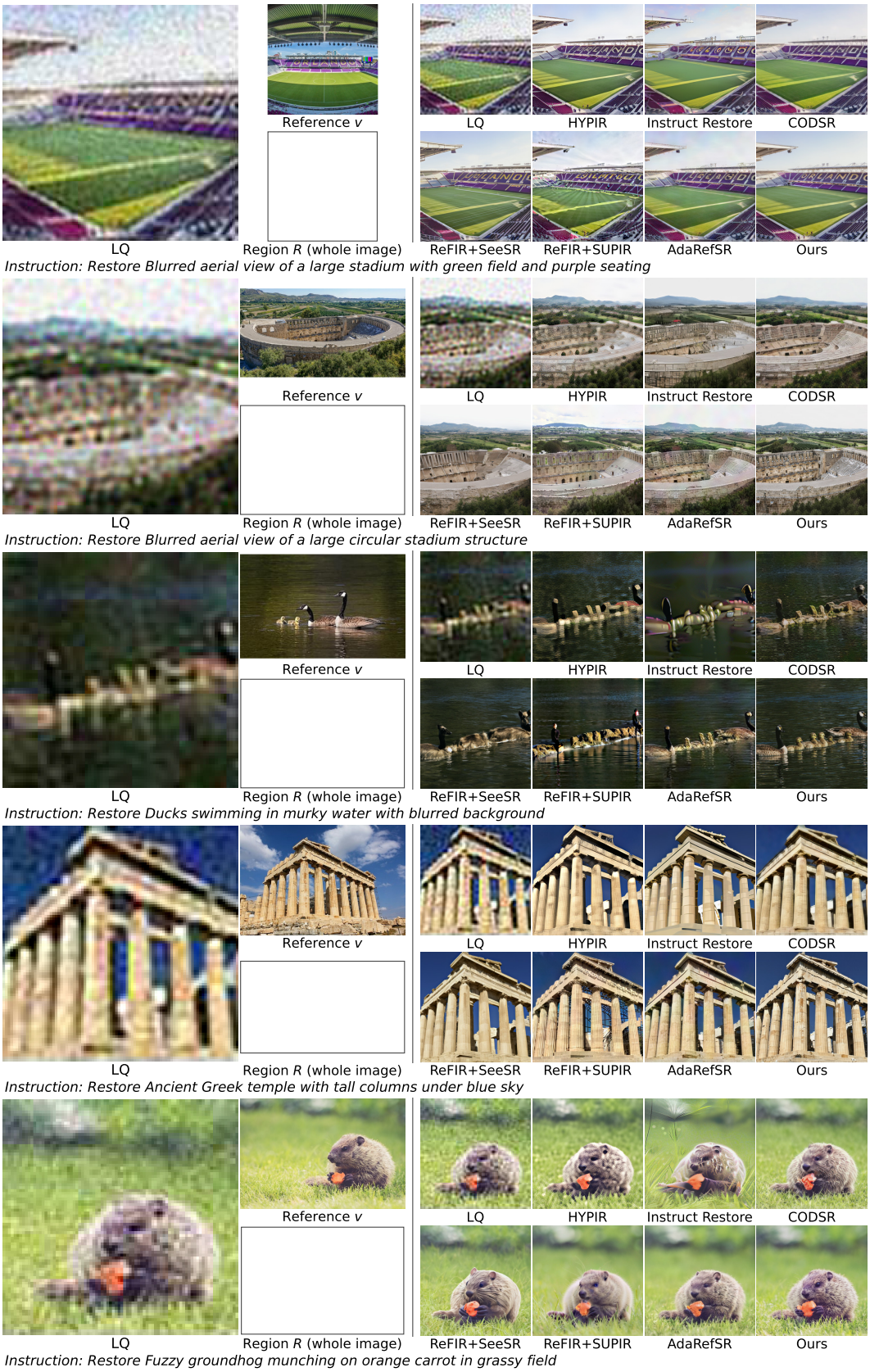

Figure 14: **Additional qualitative comparisons on WR-SR.** Each case shows the LQ input, the curated reference v, the region control R, which covers the whole image, and the textual instruction t, followed by the outputs of the compared methods.

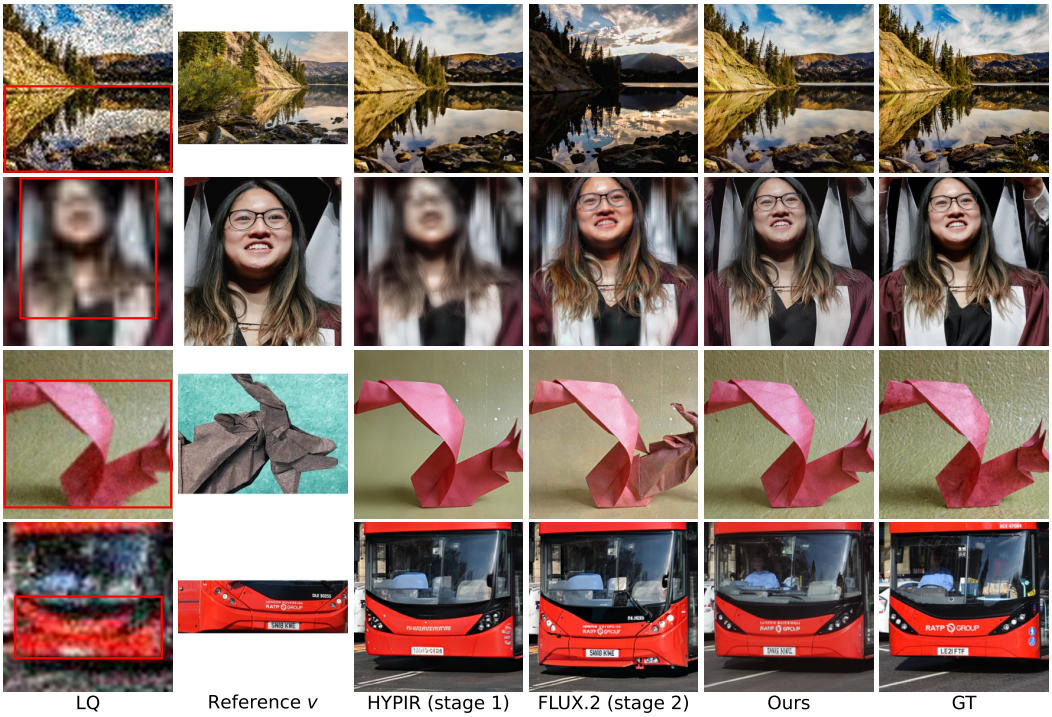

Figure 15: **FAIR compared with a two-stage cascade.** The second model receives the first model's output. In the last row, it copies the plate `SNI8 KWE` from the reference, whereas the GT reads `LE21 FTF`.

E Blinded Subjective Study: Protocol and Full Results

Platform and anonymization. The study ran on a web platform. For each trial the nine methods are rendered as an unlabeled grid; the slot-to-method assignment is re-randomized *per viewer and per trial*, and the true identities are stored only server-side, so an observer cannot infer a method from its position. Observers were not told which method was ours, nor how many methods were being compared against a proposed one.

Task and instructions. For every trial the observer selects a best and a second-best result under each of three independently presented criteria: (i) overall perceptual quality—naturalness, detail quality, and absence of over-generation; (ii) target-region correction—whether the indicated local failure was fixed and whether the result is consistent with the provided external condition; and (iii) non-target preservation—whether content outside the indicated region remained perceptually stable. The three criteria are answered separately, so a method can win one and lose another.

Participants and coverage. Twenty participants took part, each evaluating 10 cases sampled at random from our test set, for 200 evaluations per dimension and 600 judgments in total.

Statistics. Top-2 rate is the fraction of evaluations in which a method is chosen best or second-best; intervals are two-sided 95% Wilson score intervals ($z = 1.96$). Pairwise win rates are derived from the same rankings: for a given trial and dimension, a method is ranked 1, 2, or "unranked", and FAIR beats a baseline when it holds the strictly better rank. Trials in which neither method is ranked carry no ordering information and are excluded from that pair.

Table 13: **Pairwise win rates of FAIR against each baseline, derived from the blinded top-2 rankings, with two-sided 95% Wilson intervals.** FAIR wins against every baseline on all three dimensions, and every interval lower bound is at least 73.9%.

Baseline	Overall quality	Target correction	Non-target preservation
InstructRestore	99.4% (96.5–99.9)	99.4% (96.5–99.9)	99.4% (96.5–99.9)
SeeSR	95.2% (90.7–97.5)	96.3% (92.1–98.3)	95.1% (90.6–97.5)
AdaRefSR	94.6% (90.0–97.1)	93.2% (88.3–96.2)	93.4% (88.5–96.3)
ReFIR+SUPIR	93.4% (88.6–96.3)	91.6% (86.4–94.9)	92.1% (87.0–95.3)
ReFIR+SeeSR	91.1% (85.8–94.5)	91.1% (85.8–94.5)	91.1% (85.8–94.5)
CODSR	89.6% (84.2–93.3)	92.8% (87.9–95.8)	91.8% (86.7–95.1)
HYPIR	84.5% (78.6–89.1)	85.1% (79.1–89.7)	83.2% (77.1–88.0)
SUPIR	80.5% (73.9–85.7)	86.2% (80.3–90.6)	88.4% (82.7–92.3)

F Robustness of the Automatic Condition Pipeline

The optional automatic pipeline uses Qwen3-VL, SAM3, and DINOv3/CLIP to prepare restoration conditions from the low-quality input. We compare it with a GT-assisted condition oracle, smaller condition models, mismatched references, and repeated runs to assess the stability and contribution of these components.

Condition prediction accuracy. On a held-out split, the fine-tuned Qwen3-VL-8B predicts the failure type from $\mathbf{x}_{\mathrm{LQ}}$ alone with 84.03% accuracy, and 90.66% of the resulting masks reach IoU ≥ 0.5 against the annotated masks.

LQ-only inference versus a GT-assisted oracle. In the oracle setting the *only* change is that Qwen receives $(\mathbf{x}_{\mathrm{LQ}}, \mathbf{x}_{\mathrm{GT}})$ instead of $\mathbf{x}_{\mathrm{LQ}}$; SAM3 still turns the predicted region phrase into a mask, and reference acquisition and restoration are identical. Table 14 shows that the fully automatic LQ-only pipeline is close to this oracle on every metric, confirming effective GT-free operation.

Table 14: **LQ-only inference versus the GT-assisted condition oracle.** The oracle differs only in that Qwen additionally sees $\mathbf{x}_{\mathrm{GT}}$ when predicting the failure region and type; Differences are within 0.25% on all ten metrics, and within 0.12% on the no-reference and reference-similarity scores.

	PSNR↑	SSIM↑	$\mathrm{LPIPS}_{\mathrm{out}}$↓	FID↓	$\mathrm{LPIPS}_{\mathrm{in}}$↓
LQ-only (ours)	20.91	0.5582	0.2429	63.28	0.0319
GT-assisted oracle	20.91	0.5584	0.2432	63.22	0.0320
	MUSIQ↑	CLIP-IQA↑	Q-Align↑	DINO-Sim-Ref↑	CLIP-Sim-Ref↑
LQ-only (ours)	71.47	0.6728	4.514	0.4677	0.5981
GT-assisted oracle	71.50	0.6736	4.517	0.4682	0.5976

Smaller condition models. Replacing Qwen3-VL-8B with Qwen3-VL-4B, and SAM3 with SAM2.1-Hiera-Tiny ($\approx 1/15$ the parameters), leaves restoration quality essentially unchanged (Table 15); the smaller segmenter still reproduces the same regions (median mask IoU 0.873, 93.3% of regions at IoU ≥ 0.5). The gains therefore do not come from the scale of the external tools.

Table 15: **Robustness to smaller condition models.** All other settings fixed. Replacing the 8B VLM with a 4B one, or SAM3 with SAM2.1-Hiera-Tiny ($\approx 1/15$ the parameters), leaves restoration quality essentially unchanged.

Condition models	PSNR↑	SSIM↑	$\mathrm{LPIPS}_{\mathrm{out}}$↓	FID↓	$\mathrm{LPIPS}_{\mathrm{in}}$↓	MUSIQ↑	CLIP-IQA↑	Q-Align↑
Qwen3-VL-8B + SAM3 (default)	20.91	0.5582	0.2429	63.28	0.0319	71.47	0.6728	4.514
Qwen3-VL-4B + SAM3	20.91	0.5581	0.2431	63.26	0.0319	71.49	0.6728	4.514
Qwen3-VL-8B + SAM2.1-Tiny	20.92	0.5590	0.2422	62.11	0.0313	71.48	0.6726	4.518

Mismatched references. Shuffling references within each failure category (Table 16) changes PSNR, SSIM, MUSIQ, CLIP-IQA, and Q-Align by at most 0.91%, while FID, LPIPS, and reference-similarity scores show larger changes. This confirms that the model uses the reference during restoration.

Table 16: **Robustness to mismatched references.** References are shuffled within each failure category. PSNR, SSIM, MUSIQ, CLIP-IQA, and Q-Align change by at most 0.91%; FID, LPIPS, and reference-similarity scores show larger changes.

Metric	Default	Shuffled	Metric	Default	Shuffled
PSNR↑	20.91	20.80 (-0.52%)	$\text{LPIPS}_{\text{out}}$↓	0.2432	0.2490 ($+2.38\%$)
SSIM↑	0.5584	0.5533 (-0.91%)	FID↓	63.22	68.67 ($+8.63\%$)
MUSIQ↑	71.50	71.57 ($+0.10\%$)	LPIPS_{in}↓	0.0320	0.0353 ($+10.32\%$)
CLIP-IQA↑	0.6736	0.6749 ($+0.18\%$)	DINO-Sim-Ref↑	0.4682	0.4513 (-3.61%)
Q-Align↑	4.5168	4.5040 (-0.28%)	CLIP-Sim-Ref↑	0.5976	0.5884 (-1.54%)

F.1 Run-to-Run Variation

Table 17 reports mean $\pm$ standard deviation over five inference runs with different random seeds on the synthetic benchmark. Relative variation stays below 0.6% on every metric, so the margins reported in Table 1 are not attributable to run-to-run noise.

Table 17: **Mean $\pm$ standard deviation over five runs.** Relative standard deviation in parentheses.

Metric	Mean $\pm$ Std.	Metric	Mean $\pm$ Std.
PSNR	20.93 ± 0.009 (0.04%)	Q-Align	4.508 ± 0.0056 (0.12%)
SSIM	0.5586 ± 0.0002 (0.03%)	MUSIQ	71.45 ± 0.039 (0.05%)
$\text{LPIPS}_{\text{out}}$	0.2432 ± 0.0003 (0.10%)	CLIP-IQA	0.6713 ± 0.0021 (0.31%)
FID	62.84 ± 0.33 (0.52%)	DINO-Sim-Ref	0.4690 ± 0.0008 (0.16%)
LPIPS_{in}	0.0319 ± 0.00005 (0.15%)	CLIP-Sim-Ref	0.5989 ± 0.0007 (0.12%)