REFRESTORE: RESTORING ADVERSE WEATHER WITH UNALIGNED CLEAN REFERENCES

Anonymous authors
Paper under double-blind review



(a) Unaligned Clean Reference (b) Degraded Observed Input (c) **RefRestore**

Figure 1: **Reference-based restoration of real adverse weather.** Dense fog and haze destroy most of the appearance in the degraded observations (b), while the unaligned clean references (a) show the same scenes from different viewpoints, with no spatial correspondence to the target. RefRestore (c) restores the lost appearance from the reference while preserving the observed scene.

ABSTRACT

Severe adverse weather can remove the scene evidence needed for reliable restoration. Existing restoration models rely mainly on the degraded observation and learned priors; when the observation is weak, these priors cannot reliably recover details specific to the actual scene. A clean image of the same scene provides additional scene-specific information, but it is typically unaligned and may differ in viewpoint, framing, or appearance. Naively conditioning on such a reference can therefore copy content to incorrect locations or overwrite content supported by the degraded input. We introduce RefRestore, a reference-based restoration framework that exploits an unaligned same-scene image as supplementary visual information without treating it as a spatial template. RefRestore combines Content-Addressed Reference Injection, which removes explicit positional encoding from reference tokens so they are accessed by content, with Evidence-Gated Routing, which suppresses unreliable degraded features and allows other information sources to contribute where observation evidence is weak. We further introduce RefWeatherTrain-456 and RefWeatherTest-20, together with diagnostics for reference use and misplacement. On RefWeatherTest-20, RefRestore ranks first on five of six no-reference quality metrics and second on the remaining metric, while substantially reducing reference misplacement compared with naive reference conditioning. These results demonstrate the value of unaligned same-scene references as scene-specific information for adverse-weather restoration.

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.

1 INTRODUCTION

Severe adverse weather can remove or obscure the scene evidence needed for reliable image restoration. Under dense fog, snow, or rain, important appearance and structural cues become weak or unavailable. When sufficient evidence remains, a restoration model can recover much of the scene from the input itself; otherwise it must rely increasingly on learned priors to infer missing content.

Most existing adverse-weather restoration methods operate only on the degraded image. Discriminative models (Li et al., 2020; Zamir et al., 2022; Valanarasu et al., 2022; Chen et al., 2022; Patil et al., 2023) learn strong mappings from degraded to clean images, recent diffusion and generative approaches (Özdenizci & Legenstein, 2022; Chen et al., 2024; Wang et al., 2025) exploit stronger learned priors to recover perceptually plausible details, and frontier multimodal image models, including GPT-Image-2.0, Nano Banana 2, and Qwen-Image-Edit (Wu et al., 2025), demonstrate how much missing appearance can be synthesized from powerful generative priors. Nevertheless, all of these approaches face the same constraint: their priors are learned from broad image distributions, so once the observation no longer contains sufficient evidence, nothing provides information specific to the particular scene being restored, and improving the prior alone does not resolve this.

A natural source of such information is another clean photograph of the same scene. Images of the same location are often available from webcam archives, public image collections, or repeated captures, and can provide materials, colors, textures, and scene details that are not recoverable from the degraded observation alone. These images, however, are typically not pixel aligned with the target: they may differ in viewpoint, framing, illumination, time, or transient content. Consequently, naive reference conditioning can associate content with the wrong spatial locations, overwrite structures already supported by the degraded observation, or fail to exploit useful reference information when the observation is weak. The central problem we address is therefore how to exploit scene-specific information from an unaligned reference without requiring explicit pixel correspondence or treating the reference as a spatial template.

We propose RefRestore, a reference-based adverse-weather restoration framework designed for this setting. RefRestore treats the degraded image as the primary observation and the unaligned reference as a supplementary source of scene-specific appearance information. It modifies joint attention through two main mechanisms. Content-Addressed Reference Injection (CARI) removes explicit positional encoding from reference tokens, reducing dependence on their original grid locations and allowing reference features to be accessed according to content. Evidence-Gated Routing (EGR) estimates local evidence from the degraded stream and suppresses unreliable degraded keys in proportion to their estimated severity, allowing attention to redistribute toward other available information sources, including the reference, where the observation is weak. A severity-gated perceptual regularizer further emphasizes reconstruction in poorly observed regions, while mismatch-calibrated training with reference dropout and random-scene references encourages the model to fall back on the degraded observation and learned prior when the supplied reference is unreliable. Together, these components provide a practical way to incorporate unaligned scene-specific information into a strong generative restorer without requiring explicit geometric alignment (Figure 1).

The main contributions of this work are summarized as follows:

- We formulate adverse-weather restoration with unaligned same-scene references as an additional source of scene-specific visual information. To support this setting, we introduce RefWeatherTrain-456, a reference-paired training set, and RefWeatherTest-20, a real-weather benchmark with unaligned clean references, together with reference-use and reference-misplacement diagnostics that do not require pixel alignment.
- We introduce CARI and EGR as two complementary mechanisms for incorporating unaligned references into a generative restoration model, supported by severity-gated regularization and mismatch-calibrated training for selective reference use.
- Experiments on real adverse-weather scenes show that RefRestore achieves the best performance on five of six no-reference quality metrics and the second-best on the remaining metric, while substantially reducing reference misplacement compared with naive reference conditioning, demonstrating that unaligned same-scene references provide useful scene-specific information when their influence is explicitly controlled.

2 Related Work

Adverse-weather restoration has progressed from specialized methods for deraining (Yang et al., 2017; Li et al., 2019; Wang et al., 2019), desnowing (Liu et al., 2018; Chen et al., 2021; Mao et al., 2023), and dehazing (Kaiming He et al., 2009; Du & Li, 2018; Fang et al., 2024; Lan et al., 2025), to all-in-one models based on architecture search (Li et al., 2020), distillation (Chen et al., 2022), curriculum learning (Zhu et al., 2023), Transformers (Valanarasu et al., 2022; Zamir et al., 2022; Sun et al., 2024), diffusion models (Özdenizci & Legenstein, 2022; Shen et al., 2024; Xiong et al., 2025), vision-language guidance (Luo et al., 2023; Yang et al., 2023; Chen et al., 2024; Wen et al., 2025), and degradation-aware modeling (Chen & Pei, 2024; Yi et al., 2025; Wang et al., 2025; Rajagopalan & Patel, 2025). Most of these approaches still infer the restored scene primarily from the degraded observation and learned priors. Under severe degradation, however, scene-specific evidence can become weak or missing, limiting recovery of details that are not supported by the input itself. This motivates introducing an additional source of information about the same scene.

Reference-based restoration provides such external information, but existing methods make different assumptions about how the reference relates to the target. Face restoration typically relies on identity-matched references (Li et al., 2022; Hsiao et al., 2024; Zhou et al., 2025), while reference-based super-resolution uses explicit correspondence or alignment (Cao et al., 2022; Lee et al., 2025), matching-free latent priors (Liu et al., 2021), or reference images only for degradation or scale statistics (Do et al., 2025); ReBaIR (Bernasconi et al., 2025) further introduces reference information into a frozen restoration backbone. In contrast, we study adverse-weather restoration with an *unaligned image of the same scene* as supplementary scene-specific information. RefRestore is designed to use this information without requiring pixel alignment: CARI reduces explicit dependence on the reference grid location, while EGR reduces reliance on degraded features estimated to carry weak evidence. Together, they address how useful reference information can be incorporated when viewpoint correspondence is unavailable.

3 Proposed Method

Overview. RefRestore incorporates scene-specific visual information from an unaligned clean reference without treating it as a spatial template. As illustrated in Figure 2, the degraded image remains the primary observation while the reference supplies supplementary appearance information where degradation has weakened the input. The central challenge is that the two images differ in viewpoint and framing, so their grid coordinates are not comparable. Two attention mechanisms address this: Content-Addressed Reference Injection (CARI) removes explicit positional encoding from reference tokens, eliminating the bias toward reference features at coincident grid locations, and Evidence-Gated Routing (EGR) suppresses unreliable degraded keys according to a relative evidence score from the degraded stream, letting attention redistribute to other sources, including the reference, where observations are weak. An optional Region-Consistency Quarantine (see Appendix A.2 excludes reference content from user-specified semantic regions such as sky, mismatch-calibrated training with reference dropout and random-scene references keeps the model usable when the reference is uninformative, and severity-gated perceptual regularization emphasizes reconstruction in poorly observed regions.

3.1 Content-Addressed Reference Injection

Let x_0 be the clean image latent, $\epsilon \sim \mathcal{N}(0, I)$ the noise, and $\sigma_t \in [0, 1]$ the flow schedule. The noised latent is $x_t = (1 - \sigma_t)x_0 + \sigma_t\epsilon$, and the transformer predicts the velocity $v_\theta \approx v = \epsilon - x_0$ under the flow-matching objective $\mathbb{E}\,\|v_\theta - v\|^2$. In each transformer block, the tokens of the text, noised, degraded, and reference streams attend jointly,

$$A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}} + M\right), \tag{1}$$

where Q and K are the query and key matrices of the concatenated streams, d is the head dimension, and M is an additive mask, zero elsewhere, constructed by the mechanisms below. Image-side queries and keys additionally carry rotary codes for their stream and grid coordinates. For an unrotated query vector q_i at grid position p_i and an unrotated key vector k_j at position p_j, with $R(p)$

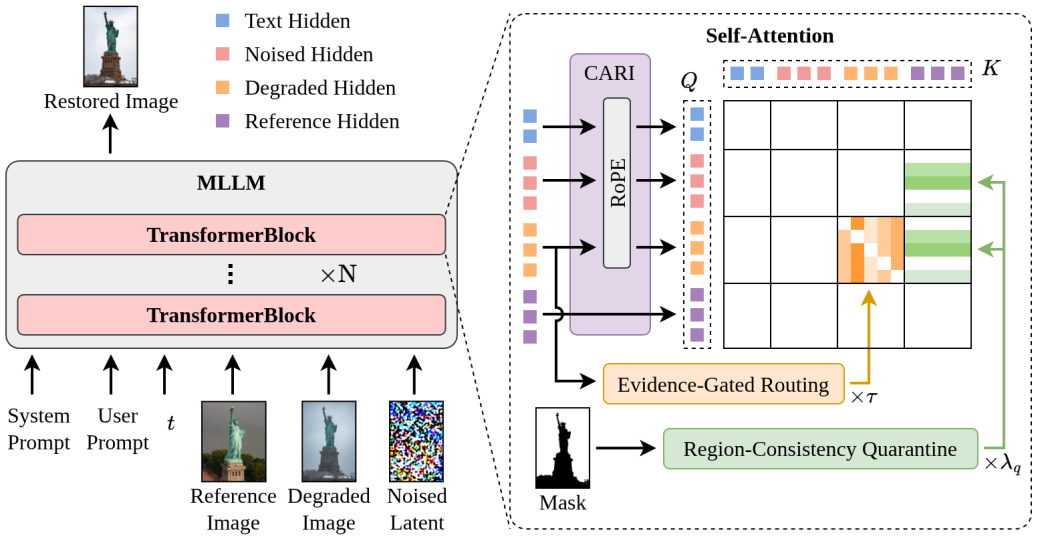


Figure 2: **Overview of RefRestore.** The degraded image, an unaligned clean reference, and the noised latent enter the backbone as separate image streams conditioned on the restoration prompt. In each joint attention block, Content-Addressed Reference Injection (CARI) removes explicit rotary position encoding from the reference tokens, reducing dependence on their original grid locations; Evidence-Gated Routing (EGR) suppresses degraded keys in proportion to a severity map estimated from the degraded stream, while preserving the diagonal; and optionally, a quarantine mask suppresses reference keys at protected query positions. Together, these mechanisms retain the degraded observation as the primary structural source while allowing supplementary appearance information from the unaligned reference to contribute where useful.

denoting the rotary transform at position p, the attention logit is

$$\ell_{ij} = \langle R(p_i)\, q_i,\ R(p_j)\, k_j \rangle = \langle q_i,\ R(p_j - p_i)\, k_j \rangle, \tag{2}$$

which introduces an explicit dependence on the relative grid offset $p_j - p_i$. This positional information is useful within the noised and degraded streams because their grids describe the target layout. For an unaligned reference, however, p_j belongs to a different image coordinate system and does not represent a valid geometric correspondence to the target. Consequently, the relative-position term can bias cross-image attention toward reference tokens according to their grid locations rather than solely according to their visual compatibility, while relevant reference content appearing elsewhere can be disadvantaged. Figure 4 illustrates the resulting content-copying failures under naive reference injection.

Content-Addressed Reference Injection (CARI) removes this explicit reference-coordinate dependence. After the per-stream rotations are computed, we restore the unrotated query and key slices on the reference tokens, leaving all projections, values, and the remaining streams unchanged. With the key-side rotation removed, the cross-stream logit becomes

$$\ell_{ij} = \langle R(p_i)\, q_i,\ k_j \rangle, \tag{3}$$

which contains no explicit term depending on the reference position p_j. Thus, reference keys are compared without encoding where they occurred in the reference grid, reducing the tendency to interpret coincident coordinates across two unaligned views as correspondence. The remaining query-side rotation is fixed across all reference keys for a given target query and therefore does not encode the reference token's location. Reference queries are likewise kept unrotated. CARI introduces no additional parameters and preserves the original joint-attention formulation; it changes how reference information is addressed, rather than introducing an explicit geometric alignment module.

3.2 Evidence-Gated Routing

Evidence estimation. Removing reference-coordinate bias does not determine how strongly the model should rely on the degraded observation itself. Under severe degradation, some regions retain

clear edges, textures, or materials, while others contain substantially weaker visual evidence. RefRestore therefore uses a lightweight feature statistic as a relative cue for the reliability of degraded tokens.

For smooth atmospheric degradations such as fog and haze, weakly observed regions tend to exhibit more homogeneous feature responses, whereas regions containing informative structure tend to activate feature channels more unevenly. We therefore use channel variance as a simple relative evidence proxy. For a degraded token i with hidden state $h_i \in \mathbb{R}^C$ at the current transformer block, where C is the channel dimension of the hidden state, we compute the variance across channels and its per-image min-max normalization,

$$v_i = \mathrm{Var}_c\left(h_{i,c}\right), \qquad \hat{e}_i = \mathrm{clamp}_{[0,1]}\left(\frac{v_i - \min_j v_j}{\max_j v_j - \min_j v_j + \varepsilon}\right), \tag{4}$$

where $h_{i,c}$ is the c-th channel of h_i, v_i is its variance over the C channels, and the minimum and maximum are taken over all degraded tokens j of the same image. The normalized value $\hat{e}_i \in [0,1]$ provides a relative ranking within each image rather than a calibrated uncertainty estimate. The constant ε guards against a degenerate variance range; in this case we set $\hat{e}_i = 1$, corresponding to no additional suppression. The evidence rank is converted into degradation severity as

$$s_i = (1 - \hat{e}_i)^\gamma, \tag{5}$$

where $\gamma = 2$ sharpens the ranking. The severity map s is recomputed at every block and timestep from the current hidden states, allowing the routing cue to evolve with the restoration process. Larger s_i indicates that token i is treated as less reliable under this relative evidence heuristic. We emphasize that this statistic is a routing cue rather than a general uncertainty estimator; its limitations for high-frequency degradations such as rain and snow are discussed in Appendix A.3.

Routing. Given the severity map, we apply an additive penalty to the degraded-to-degraded attention block,

$$M_{ij} = -\tau\, s_j\,(1 - \delta_{ij}) \quad \text{for } i, j \in \mathcal{D}, \qquad A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}} + M\right), \tag{6}$$

zero elsewhere, where $\mathcal{D}$ indexes the degraded tokens, i and j index degraded query and key tokens, d is the head dimension, $\tau \geq 0$ is the evidence strength and sole routing hyperparameter, s_j is the estimated severity of degraded key j, and δ_{ij} spares the diagonal.

The key-side penalty reduces attention to degraded features that are estimated to carry weak evidence. Because this penalty is applied before the softmax, the released attention mass can redistribute among the remaining information sources, including reliable degraded tokens, the noised stream, text conditioning, and the position-independent reference tokens provided by CARI. Importantly, EGR does not explicitly force a query to attend to the reference; instead, it reduces the dominance of unreliable degraded context and allows other available sources to contribute when useful. Sparing the diagonal preserves direct access to each token's own spatially aligned observation, while suppressing unreliable neighbouring keys reduces the tendency of strongly degraded regions to reinforce one another through self-attention.

CARI and EGR therefore address complementary aspects of reference-assisted restoration. CARI reduces explicit coordinate bias when accessing an unaligned reference, while EGR modulates the influence of unreliable degraded features. Their combination allows scene-specific reference information to participate in restoration without requiring explicit geometric registration. An optional region-consistency quarantine (Appendix A.2) can further restrict direct reference access in user-specified semantic regions such as sky.

3.3 Reference-Use Diagnostics

Because the reference is unaligned, its pixels cannot serve as a direct ground-truth target for the restored image. We therefore introduce two tuning-free diagnostics based on patch-level DINOv2 (Oquab et al., 2024) features of the degraded image, reference, and prediction. ***RefUse*** measures how strongly the prediction moves toward content represented in the reference, particularly where the degraded input has weak compatibility with that content. ***RefMisplace*** measures signs of

potentially inappropriate reference transfer, including reference content that appears in the prediction without support from the degraded observation and cases where the prediction becomes more reference-like than observation-like.

Let D_i, R_j, and P_i be the DINO patch tokens of the degraded image, the reference, and the prediction, resized to the degraded grid, with $i = 1, \ldots, N$ over the target grid and $j = 1, \ldots, N_r$ over the reference grid, and let $S^{dr}_{ij} = \cos(D_i, R_j)$ and $S^{pr}_{ij} = \cos(P_i, R_j)$.

RefUse. RefUse measures the increase in similarity between the prediction and its nearest reference content relative to the degraded input, with greater weight assigned to target patches that initially have low compatibility with the reference. We define $\mathrm{compat}_i = \max(0, \max_j S^{dr}_{ij})$, where a smaller value indicates that degraded patch i has no strong match among the reference patches.

$$\mathrm{RefUse} = \frac{\sum_i w_i \,\mathrm{relu}\big(\max_j S^{pr}_{ij} - \mathrm{compat}_i\big)}{\sum_i w_i}, \qquad w_i = 1 - \mathrm{compat}_i, \tag{7}$$

The ReLU term is positive only when the prediction becomes more similar to some reference content than the degraded input was, while w_i gives larger weight to patches with weaker initial compatibility. A higher RefUse therefore indicates a stronger shift of the restored representation toward content available in the reference.

RefMisplace. RefMisplace captures two complementary indicators of potentially inappropriate reference transfer. The *intrusion* term identifies reference content that has weak support in the degraded observation but appears strongly in the prediction. For reference patch j, we define $a_j = \mathrm{clamp}_{[0,1]}(1 - \max_i S^{dr}_{ij})$, which is large when no degraded patch strongly matches reference patch j, and $o_j = \mathrm{clamp}_{[0,1]}(\max_i S^{pr}_{ij})$, which is large when some prediction patch strongly matches it. The *overwrite* term measures target patches for which the prediction becomes more similar to reference content than to its own degraded observation, using $\mathrm{selfsim}_i = \cos(P_i, D_i)$:

$$\mathrm{RefMisplace} = \underbrace{\frac{1}{N_r}\sum_j a_j\, o_j}_{\text{intrusion}} + \underbrace{\frac{1}{N}\sum_i \mathrm{relu}\left(\max_j S^{pr}_{ij} - \mathrm{selfsim}_i\right)}_{\text{overwrite}}. \tag{8}$$

A larger RefMisplace indicates stronger evidence, under the DINO representation, that reference-related content may have been introduced without corresponding support from the degraded observation or may have overwritten observable content.

Interpretation. Both quantities are diagnostics of reference-related behavior, not measures of factual correctness or causal reference usage: a good reference-free restoration may also move toward a clean reference, severe degradation may hide genuine content from the degraded features, and DINO features capture semantic content rather than exact geometry, so shape changes within a semantic class can escape detection. We therefore interpret them together with qualitative inspection, and sanity-check them on examples with known reference-copying behavior in Appendix A.3.

4 Experiments

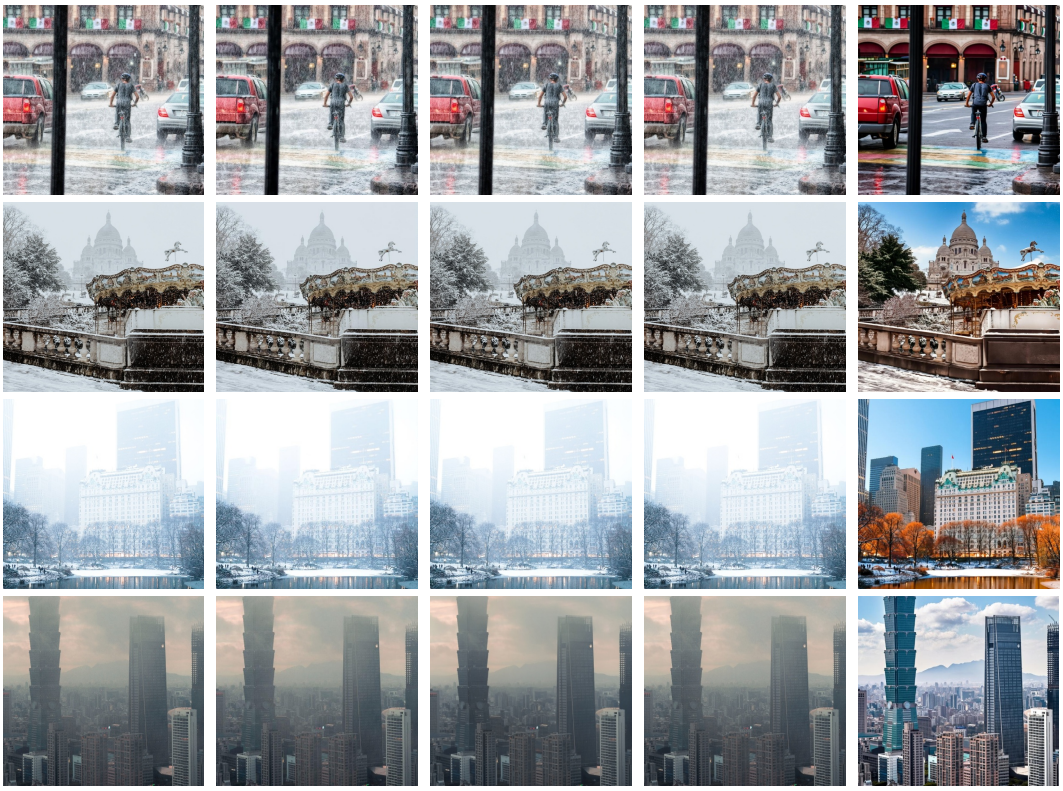
We implement RefRestore on Qwen-Image-Edit, whose joint attention over text and multiple image streams naturally accommodates the noised, degraded, and reference tokens, while its per-stream rotary embeddings provide the positional mechanism modified by CARI. The released backbone is otherwise retained, with adaptation focused on the proposed attention mechanisms and training strategy. We train on RefWeatherTrain-456, which contains 456 images from 52 scenes; images from the same scene share salient scene content while differing in viewpoint and camera settings. Detailed training settings, including adaptation rank, optimizer, schedule, dropout probabilities, and loss weights, are provided in Appendix A.1.

4.1 Evaluation

We evaluate on RefWeatherTest-20, which contains 20 held-out real weather scenes, each paired with an unaligned clean image of the same scene captured under diverse, unknown camera and

Table 1: **Comparison RefWeatherTest-20 benchmark.** Our method achieves the best TReS, NIQE, MUSIQ, CLIPIQA+ and Q-Align scores. **Bold** and underline denote best and second-best results, respectively. Arrows ↑ and ↓ indicate higher and lower are better.

Method	TReS↑	NIQE↓	HyperIQA↑	MUSIQ↑	CLIPIQA+↑	Q-Align↑
TKL-MCR (2022)	67.53	3.46	0.436	58.03	0.543	3.75
WeatherDiff (2022)	70.41	3.46	0.499	61.75	0.591	3.81
MWDT (2023)	63.39	4.19	0.515	58.12	0.532	3.69
T^3-DiffWeather (2024)	69.07	3.48	0.492	61.49	0.582	3.83
SSGFormer (2025)	66.61	3.32	0.495	60.83	0.587	3.88
MODEM (2025)	67.99	4.64	**0.573**	61.90	0.608	3.87
Qwen-Image-Edit (2025)	66.85	3.11	0.492	66.61	0.655	4.32
RefRestore	**78.22**	**2.77**	0.558	**71.05**	**0.712**	**4.59**



(a) Input (b) T^3-DiffWeather (c) SSGFormer (d) MODEM (e) **RefRestore**

Figure 3: **Qualitative comparison on RefWeatherTest-20 benchmark.** Where state-of-the-art baselines struggle on real scenes, leaving residual haze, snow, and washed-out color, RefRestore uses an unaligned clean reference to produce clear and vivid results.

processing conditions; no test scene appears in RefWeatherTrain-456. Around this benchmark we address two questions: can an unaligned same-scene reference provide useful scene-specific information for restoring real adverse-weather images, and can this information be incorporated without the reference-copying behavior of naive conditioning? Because pixel-aligned clean ground truth is unavailable, we assess the first through no-reference perceptual-quality metrics and qualitative comparisons, and the second through controlled component ablations with the RefUse and RefMisplace diagnostics of Section 3. The quality metrics measure whether the output looks natural and clear, while the diagnostics characterize reference-associated changes without establishing factual correctness.

Table 2: **Ablation of the evidence strength τ on RefWeatherTest-20.** Parenthesized values are deviations from the no-reference baseline (first row).

Method	τ	RefUse↑	RefMisplace↓
Baseline (w/o. reference)	–	0.0545	0.2357
Baseline (w/. reference)	–	0.1748 (+0.1203)	0.5471 (+0.3114)
RefRestore	0.1	0.0661 (+0.0116)	0.2458 (+0.0101)
RefRestore	1.0	0.0648 (+0.0103)	0.2449 (+0.0092)
RefRestore	3.0	0.0580 (+0.0035)	0.2362 (+0.0005)

Table 3: **Incremental component ablation on RefWeatherTest-20.** Each row adds one component to the previous configuration. Detailed discussion in Section 4.2.

Method	TReS↑	NIQE↓	HyperIQA↑	MUSIQ↑	CLIPIQA+↑	Q-Align↑	RefUse↑	RefMisplace↑
Baseline	66.85	3.11	0.492	66.61	0.655	4.32	0.0545	0.236
+Reference	74.72	2.92	0.552	70.56	0.724	4.68	0.1748	0.547
+Finetune	79.96	2.59	0.566	72.14	0.727	4.68	0.1839	0.558
+CARI-EGR	75.54	2.69	0.531	70.26	0.699	4.54	0.0667	0.249
+SGR	78.22	2.77	0.558	71.05	0.712	4.59	0.0648	0.245

We compare RefRestore against representative adverse-weather restoration methods, TKL-MCR (Chen et al., 2022), WeatherDiff (Özdenizci & Legenstein, 2022), MWDT (Patil et al., 2023), T^3-DiffWeather (Chen et al., 2024), SSGFormer (Jeong et al., 2025), and MODEM (Wang et al., 2025), as well as the released Qwen-Image-Edit backbone (Wu et al., 2025). Unless otherwise stated, RefRestore uses evidence strength $\tau = 1.0$. The optional region-consistency quarantine in Appendix A.2 is disabled so evaluation isolates the automatic reference-use mechanisms without user-provided semantic masks. Full qualitative results are provided in Appendix A.5.

Quantitative Results. Table 1 reports performance on six complementary no-reference quality metrics: TReS (Mittal et al., 2012), NIQE (Mittal et al., 2013), HyperIQA (Su et al., 2020), MUSIQ (Ke et al., 2021), CLIPIQA+ (Wang et al., 2022), and Q-Align (Wu et al., 2023). RefRestore ranks first on five of the six metrics and second on HyperIQA. In particular, it improves substantially over the released Qwen-Image-Edit backbone on TReS, MUSIQ, CLIPIQA+, and Q-Align, indicating that the complete reference-assisted training and inference framework produces higher perceptual quality than relying on the pretrained generative prior alone. Different metrics do not rank all methods identically, as expected for no-reference image-quality assessment. Nevertheless, RefRestore performs consistently strongly across both conventional and learned perceptual measures. These results establish the perceptual benefit of the full system, while the ablations in Section 4.2 examine how the reference and individual components contribute to this improvement.

Qualitative Results. Figure 3 compares representative real-weather restorations. The reference-free baselines often retain substantial atmospheric veil, precipitation, or washed-out appearance in severely degraded regions. RefRestore produces visibly clearer outputs with more defined distant structures and stronger material and color appearance. Importantly, because RefWeatherTest-20 does not provide pixel-aligned clean ground truth, these examples demonstrate perceptual restoration quality rather than factual recovery of every reconstructed detail. The reference-use behavior of RefRestore is therefore examined separately through the controlled comparisons in Section 4.2.

4.2 Ablation Studies

Our central question is whether reference information can improve restoration without dominating content supported by the degraded observation. We therefore construct an incremental ablation from the released backbone to the full system, separating the benefit of providing a reference from the mechanisms that control its influence.



(a) Reference (b) Input (c) Baseline (d) +Reference (e) +Finetune (f) +CARI-EGR (g) +SGR

Figure 4: **Incremental component ablation.** Outputs of the configurations in Table 3 on two RefWeatherTest-20 scenes. Each column adds one component to the baseline (c).

Contribution of individual components. Table 3 reveals the trade-off that motivates RefRestore. Simply supplying the reference (+Reference) improves all perceptual metrics, showing that same-scene information can be useful, but raises RefMisplace from 0.236 to 0.547. Finetuning (+Finetune) sharpens further while misuse remains high (0.558), so better-looking restoration alone does not imply appropriate incorporation. Adding CARI and EGR drops RefMisplace to 0.249, near the no-reference floor, and removes content copied under naive conditioning, such as the crowd and graffiti bin in Figure 4 (row 1). This reduction comes at some perceptual cost, consistent with a more conservative use of additional reference information. Adding the severity-gated regularizer (+SGR) then recovers much of that perceptual-quality gap (TReS $75.54 \rightarrow 78.22$, HyperIQA $0.531 \rightarrow 0.558$) while keeping RefMisplace low at 0.245, confirming the intended roles of the components: information from reference, transfer constrained by CARI and EGR, and quality restored by SGR under controlled reference use. We interpret RefUse more cautiously. It rises strongly under naive reference conditioning and remains only modestly above the no-reference baseline for RefRestore. As discussed in Section 3, RefUse measures movement toward reference content rather than causal usage. The combination of the quantitative diagnostics and the controlled qualitative comparisons therefore provides stronger evidence than either diagnostic alone.

Evidence strength for routing. The routing strength τ controls how strongly degraded keys with high estimated severity are suppressed. Table 2 shows that naive reference conditioning produces the largest increase in both RefUse and RefMisplace: the output moves substantially toward reference appearance, but the reference-misplacement diagnostic also increases from 0.2357 to 0.5471. RefRestore exhibits substantially more restrained behavior. Across $\tau \in \{0.1, 1.0, 3.0\}$, RefMisplace remains within 0.0101 of the no-reference baseline, while RefUse remains slightly above it. Increasing τ reduces both deviations from the no-reference baseline. This is consistent with progressively stronger suppression of degraded contextual keys producing more conservative reference-associated changes in the final output. We use $\tau = 1.0$ as the default setting because it retains the low RefMisplace behavior while providing the restoration performance reported in Table 1. Since RefUse and RefMisplace are diagnostics rather than calibrated quantities on a shared scale, we do not interpret their ratio; instead, we use them together with the qualitative ablations to characterize how reference conditioning changes as routing strength varies.

5 Conclusion

We presented RefRestore, a reference-based framework for adverse-weather restoration that uses an unaligned clean image of the same scene as supplementary scene-specific information. Content-Addressed Reference Injection reduces explicit reference-coordinate bias, while Evidence-Gated Routing suppresses unreliable degraded features so other information sources can contribute where the observation is weak. Severity-gated regularization and mismatch-calibrated training further support controlled reference use. Experiments show strong perceptual performance while substantially reducing reference misplacement compared with naive conditioning, demonstrating the potential of unaligned same-scene references for severe real-world restoration.

AI USE STATEMENT

In this work, we used generative AI tools for [design and provide feedback on research methodology or experiments, implement methods]. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks, generate synthetic data sets, formulate mathematical claims, clean and reformat dataset, support qualitative and thematic data analysis, interpret results] and [provide critical ingredients for proving mathematical claims, assist with translation] are not applicable to this work. Additionally, we used generative AI tools for [create or edit software code, creation of artifacts, draft parts of a research paper, summarize or analyse existing literature, discover research topics or identify gaps, brainstorming, sourcing/searching for information, edit a research paper to improve readability, suggest a structure for a research paper, propose a title or keywords for a research paper]. We have reviewed all AI-assisted work. [We perform a manual literature survey on relevant work before using LLM for summary, all LLM-generated code was verified and tested for correctness, we review all LLM-generated outputs]. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Michael Bernasconi, Abdelaziz Djelouah, Yang Zhang, Markus Gross, and Christopher Schroers. ReBaIR: Reference-Based Image Restoration. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 5548–5557, Honolulu, HI, USA, October 2025. IEEE. ISBN 9798331589882. doi: 10.1109/ICCVW69036.2025.00581. URL https://ieeexplore.ieee.org/document/11375704/.

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, 2022. URL https://arxiv.org/abs/2207.11938.

Sixiang Chen, Tian Ye, Kai Zhang, Zhaohu Xing, Yunlong Lin, and Lei Zhu. Teaching Tailored to Talent: Adverse Weather Restoration via Prompt Pool and Depth-Anything Constraint, 2024. URL https://arxiv.org/abs/2409.15739.

Wei-Ting Chen, Hao-Yu Fang, Cheng-Lin Hsieh, Cheng-Che Tsai, I-Hsiang Chen, Jian-Jiun Ding, and Sy-Yen Kuo. ALL Snow Removed: Single Image Desnowing Algorithm Using Hierarchical Dual-tree Complex Wavelet Representation and Contradict Channel Loss. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4176–4185, Montreal, QC, Canada, October 2021. IEEE. ISBN 9781665428125. doi: 10.1109/ICCV48922.2021.00416. URL https://ieeexplore.ieee.org/document/9710202/.

Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen Kuo. Learning Multiple Adverse Weather Removal via Two-stage Knowledge Learning and Multi-contrastive Regularization: Toward a Unified Model. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17632–17641, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-66546-946-3. doi: 10.1109/CVPR52688.2022.01713. URL https://ieeexplore.ieee.org/document/9879902/.

Yu-Wei Chen and Soo-Chang Pei. Always Clear Days: Degradation Type and Severity Aware All-In-One Adverse Weather Removal, March 2024. URL http://arxiv.org/abs/2310.18293. arXiv:2310.18293 [cs].

Huu-Phu Do, Po-Chih Hu, Hao-Chien Hsueh, Che-Kai Liu, Vu-Hoang Tran, and Ching-Chun Huang. Blind Super Resolution with Reference Images and Implicit Degradation Representation, 2025. URL https://arxiv.org/abs/2507.13915.

Yixin Du and Xin Li. Perceptually Optimized Generative Adversarial Network for Single Image Dehazing, May 2018. URL http://arxiv.org/abs/1805.01084. arXiv:1805.01084 [cs].

Chengyu Fang, Chunming He, Fengyang Xiao, Yulun Zhang, Longxiang Tang, Yuelin Zhang, Kai Li, and Xiu Li. Real-world Image Dehazing with Coherence-based Label Generator and Cooperative Unfolding Network, October 2024. URL http://arxiv.org/abs/2406.07966. arXiv:2406.07966 [cs].

Qing Guo, Jingyang Sun, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Wei Feng, and Yang Liu. EfficientDeRain: Learning Pixel-wise Dilation Filtering for High-Efficiency Single-Image Deraining, September 2020. URL http://arxiv.org/abs/2009.09238. arXiv:2009.09238 [cs].

Chi-Wei Hsiao, Yu-Lun Liu, Cheng-Kun Yang, Sheng-Po Kuo, Kevin Jou, and Chia-Ping Chen. ReF-LDM: A Latent Diffusion Model for Reference-based Face Image Restoration, 2024. URL https://arxiv.org/abs/2412.05043.

Yuhwan Jeong, Yunseo Yang, Youngho Yoon, and Kuk-Jin Yoon. Robust Adverse Weather Removal via Spectral-based Spatial Grouping, 2025. URL https://arxiv.org/abs/2507.22498.

Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1956–1963, Miami, FL, June 2009. IEEE. ISBN 978-1-4244-3992-8. doi: 10.1109/CVPR.2009.5206515. URL https://ieeexplore.ieee.org/document/5206515/.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale Image Quality Transformer, August 2021. URL http://arxiv.org/abs/2108.05997. arXiv:2108.05997 version: 1.

Yunwei Lan, Zhigao Cui, Xin Luo, Chang Liu, Nian Wang, Menglin Zhang, Yanzhao Su, and Dong Liu. When Schrödinger Bridge Meets Real-World Image Dehazing with Unpaired Training, July 2025. URL http://arxiv.org/abs/2507.09524. arXiv:2507.09524.

Byeonghun Lee, Hyunmin Cho, Hong Gyu Choi, Soo Min Kang, Iljun Ahn, and Kyong Hwan Jin. Reference-Based Super-Resolution via Image-Based Retrieval-Augmented Generation Diffusion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10764–10774, Honolulu, HI, USA, October 2025. IEEE. ISBN 9798331587758. doi: 10.1109/ICCV51701.2025.01002. URL https://ieeexplore.ieee.org/document/11445011/.

Ruoteng Li, Robby T. Tan, and Loong-Fah Cheong. All in One Bad Weather Removal Using Architectural Search. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3172–3182, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.00324. URL https://ieeexplore.ieee.org/document/9157460/.

Ruotent Li, Loong Fah Cheong, and Robby T. Tan. Heavy Rain Image Restoration: Integrating Physics Model and Conditional Adversarial Learning, April 2019. URL http://arxiv.org/abs/1904.05050. arXiv:1904.05050 [cs].

Xiaoming Li, Shiguang Zhang, Shangchen Zhou, Lei Zhang, and Wangmeng Zuo. Learning Dual Memory Dictionaries for Blind Face Restoration, 2022. URL https://arxiv.org/abs/2210.08160.

Beibei Lin, Yeying Jin, Wending Yan, Wei Ye, Yuan Yuan, and Robby T. Tan. NightHaze: Nighttime Image Dehazing via Self-Prior Learning, March 2024. URL http://arxiv.org/abs/2403.07408. arXiv:2403.07408 [cs].

Yun-Fu Liu, Da-Wei Jaw, Shih-Chia Huang, and Jenq-Neng Hwang. DesnowNet: Context-Aware Deep Network for Snow Removal. *IEEE Transactions on Image Processing*, 27(6):3064–3073, June 2018. ISSN 1057-7149, 1941-0042. doi: 10.1109/TIP.2018.2806202. URL http://arxiv.org/abs/1708.04512. arXiv:1708.04512 [cs].

Zhi-Song Liu, Wan-Chi Siu, and Li-Wen Wang. Variational AutoEncoder for Reference based Image Super-Resolution. 2021. doi: 10.48550/ARXIV.2106.04090. URL https://arxiv.org/abs/2106.04090.

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Controlling Vision-Language Models for Multi-Task Image Restoration, 2023. URL https://arxiv.org/abs/2310.01018.

Jiawei Mao, Yuanqi Chang, Xuesong Yin, and Binling Nie. Star-Net: Improving Single Image Desnowing Model With More Efficient Connection and Diverse Feature Interaction, 2023. URL https://arxiv.org/abs/2303.09988.

A. Mittal, A. K. Moorthy, and A. C. Bovik. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, December 2012. ISSN 1057-7149, 1941-0042. doi: 10.1109/TIP.2012.2214050. URL http://ieeexplore.ieee.org/document/6272356/.

A. Mittal, R. Soundararajan, and A. C. Bovik. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, March 2013. ISSN 1070-9908, 1558-2361. doi: 10.1109/LSP.2012.2227726. URL http://ieeexplore.ieee.org/document/6353522/.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. URL http://arxiv.org/abs/2304.07193. arXiv:2304.07193 [cs].

Prashant W. Patil, Sunil Gupta, Santu Rana, Svetha Venkatesh, and Subrahmanyam Murala. Multi-weather Image Restoration via Domain Translation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21639–21648, Paris, France, October 2023. IEEE. ISBN 9798350307184. doi: 10.1109/ICCV51070.2023.01983. URL https://ieeexplore.ieee.org/document/10377637/.

Sudarshan Rajagopalan and Vishal M. Patel. AWRaCLe: All-Weather Image Restoration Using Visual In-Context Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (6):6675–6683, April 2025. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v39i6.32716. URL https://ojs.aaai.org/index.php/AAAI/article/view/32716.

Yiyang Shen, Mingqiang Wei, Yongzhen Wang, Xueyang Fu, and Jing Qin. Rethinking Real-world Image Deraining via An Unpaired Degradation-Conditioned Diffusion Model, May 2024. URL http://arxiv.org/abs/2301.09430. arXiv:2301.09430 [cs].

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3664–3673, Seattle, WA, USA, June 2020. IEEE. ISBN 9781728171685. doi: 10.1109/CVPR42600.2020.00372. URL https://ieeexplore.ieee.org/document/9156687/.

Shangquan Sun, Wenqi Ren, Xinwei Gao, Rui Wang, and Xiaochun Cao. Restoring Images in Adverse Weather Conditions via Histogram Transformer, 2024. URL https://arxiv.org/abs/2407.10172.

Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M. Patel. TransWeather: Transformer-based Restoration of Images Degraded by Adverse Weather Conditions, June 2022. URL http://arxiv.org/abs/2111.14813. arXiv:2111.14813 [cs].

Hainuo Wang, Qiming Hu, and Xiaojie Guo. MODEM: A Morton-Order Degradation Estimation Mechanism for Adverse Weather Image Recovery, 2025. URL https://arxiv.org/abs/2505.17581.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for Assessing the Look and Feel of Images, 2022. URL https://arxiv.org/abs/2207.12396.

Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson Lau. Spatial Attentive Single-Image Deraining with a High Quality Real Rain Dataset, August 2019. URL http://arxiv.org/abs/1904.01538. arXiv:1904.01538 [cs].

Yuanbo Wen, Tao Gao, Jing Zhang, Ziqi Li, and Ting Chen. Multi-axis Prompt and Multi-dimension Fusion Network for All-in-one Weather-degraded Image Restoration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):8323–8331, April 2025. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v39i8.32898. URL https://ojs.aaai.org/index.php/AAAI/article/view/32898.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image Technical Report, 2025. URL https://arxiv.org/abs/2508.02324.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels, 2023. URL https://arxiv.org/abs/2312.17090.

Jiamei Xiong, Xuefeng Yan, Yongzhen Wang, Wei Zhao, Xiao-Ping Zhang, and Mingqiang Wei. DA2Diff: Exploring Degradation-aware Adaptive Diffusion Priors for All-in-One Weather Restoration, 2025. URL https://arxiv.org/abs/2504.05135.

Hao Yang, Liyuan Pan, Yan Yang, and Wei Liang. Language-driven All-in-one Adverse Weather Removal, 2023. URL https://arxiv.org/abs/2312.01381. Version Number: 1.

Wenhan Yang, Robby T. Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep Joint Rain Detection and Removal from a Single Image, March 2017. URL http://arxiv.org/abs/1609.07769. arXiv:1609.07769 [cs].

Jingjun Yi, Qi Bi, Hao Zheng, Huimin Huang, Yixian Shen, Haolan Zhan, Wei Ji, Yawen Huang, Yuexiang Li, Xian Wu, and Yefeng Zheng. Degradation-Aware Dynamic Schrödinger Bridge for Unpaired Image Restoration. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=jJweE513Am.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient Transformer for High-Resolution Image Restoration, March 2022. URL http://arxiv.org/abs/2111.09881. arXiv:2111.09881 [cs].

Mo Zhou, Keren Ye, Viraj Shah, Kangfu Mei, Mauricio Delbracio, Peyman Milanfar, Vishal M. Patel, and Hossein Talebi. Reference-Guided Identity Preserving Face Restoration, 2025. URL https://arxiv.org/abs/2505.21905.

Yurui Zhu, Tianyu Wang, Xueyang Fu, Xuanyu Yang, Xin Guo, Jifeng Dai, Yu Qiao, and Xiaowei Hu. Learning Weather-General and Weather-Specific Features for Image Restoration Under Multiple Adverse Weather Conditions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21747–21758, Vancouver, BC, Canada, June 2023. IEEE. ISBN 9798350301298. doi: 10.1109/CVPR52729.2023.02083. URL https://ieeexplore.ieee.org/document/10204275/.

Ozan Özdenizci and Robert Legenstein. Restoring Vision in Adverse Weather Conditions with Patch-Based Denoising Diffusion Models, November 2022. URL http://arxiv.org/abs/2207.14626. arXiv:2207.14626 [cs].

A Appendix

A.1 Training Details

We train solely on RefWeatherTrain-456. Each training sample selects one image of a scene as the clean ground truth, synthesizes its degraded counterpart on the fly, and draws an unaligned reference from the remaining images of the same scene. Rain and snow degradations are rendered with overlays adapted from prior work (Guo et al., 2020; Liu et al., 2018; Li et al., 2019), while haze follows the severe synthesis pipeline of NightHaze (Lin et al., 2024). Reference conditioning follows the mismatch-calibrated regime: with probability 0.2 the reference is dropped entirely, and when retained it is replaced by a random-scene reference with probability 0.2, teaching a reference-free fallback and the rejection of untrustworthy references; for all training samples, the evidence strength is $\tau = 1.0$. The objective combines the flow-matching velocity loss with the severity-gated LPIPS regularizer at weight $\lambda = 0.3$, which concentrates perceptual pressure where evidence is absent. The weights for severity-gate is computed the same as in Section 3.2. We adapt only the image-stream projections with LoRA of rank 64, keeping all remaining weights, including the text interface, frozen, and optimize with AdamW (weight decay 0.01) at a learning rate of 2×10^{-4}, batch size 1 with gradient accumulation over 4 steps, for 5000 epochs on a single H200 GPU.

A.2 Region-Consistency Quarantine

Evidence-Gated Routing redirects evidence-poor queries to the reference, making sky tokens vulnerable to wrong-class transfer when the reference depicts a different sky or none. As an optional deployment-time guardrail, the quarantine penalizes reference keys for both query streams that read the reference,

$$M_{ij} \mathrel{-}= \lambda_q \, m_i \qquad \text{for } (i,j) \in (\mathcal{N} \times \mathcal{R}) \cup (\mathcal{D} \times \mathcal{R}), \tag{9}$$

where $m_i \in [0,1]$ is a soft mask giving the fraction of token i's receptive field labeled as the protected class, and λ_q is the quarantine strength, set above the largest bias routing can apply. Because the penalty is uniform across the reference columns of row i, it caps that row's total reference mass regardless of content similarity, forcing protected regions to synthesize appearance from the prior and their own tokens. The mask may come from user input, a vision-language probe, or latent statistics. Scope: the quarantine caps direct reference reads by protected rows; reference-derived content that enters through unprotected rows can still propagate via same-stream blocks, so the guardrail complements, rather than substitutes for, rejecting globally mismatched references. All reported experiments disable it, since every test reference depicts the same scene as its target.

A.3 Further Ablation Studies

Prior strength without faithfulness. Figure 5 compares RefRestore with frontier MLLMs, the strongest image priors among compared methods, on real degraded observations. Their outputs are sharp and veil-free, but the highlighted regions show restoration by re-imagination: GPT-Image-2.0 distorts the tower and removes the shrub, Nano Banana 2 deletes people and adds a car, and Qwen-Image-Edit hallucinates shadows and removes pedestrians. Because their conditioning does not separate what the observation determines from what must be synthesized, any region can be silently altered. RefRestore instead anchors existence and structure to the degraded input and borrows appearance only where evidence is absent, leaving the highlighted content unchanged.

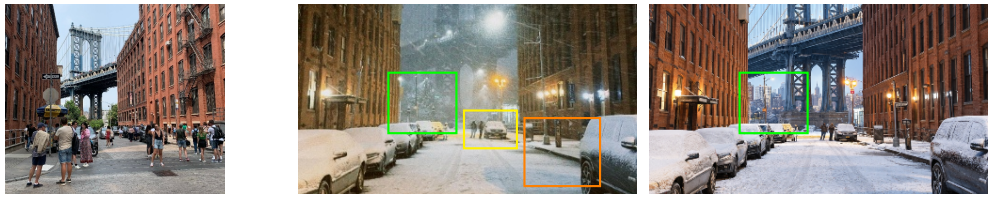
Sanity check on RefUse and RefMisplace. To verify that RefUse and RefMisplace measure what they claim, we score output pairs with a known faithfulness ordering on RefWeatherTest-20 scenes in Figure 6. Outputs in column (c) visibly copies reference content, unlike (d) which preserves it. In every scene the metric ordering matches the visual evidence: column (c) scores RefMisplace in the range of $[0.38, 0.70]$ against column (d)'s range of $[0.14, 0.31]$, and the gap widens exactly where the intrusion is most overt, $+0.32$ for summer foliage and building pasted in row 1, $+0.24$ for the copied crowd and background buildings in row 3, and $+0.41$ for duplicated cars in row 4. RefUse is likewise higher for naive injection ($[0.05, 0.30]$ versus $[0.02, 0.11]$), reflecting indiscriminate borrowing rather than selective transfer. Because both outputs are scored against the same triplet with no tuned thresholds, this consistent per-scene separation confirms that the metrics respond to con-



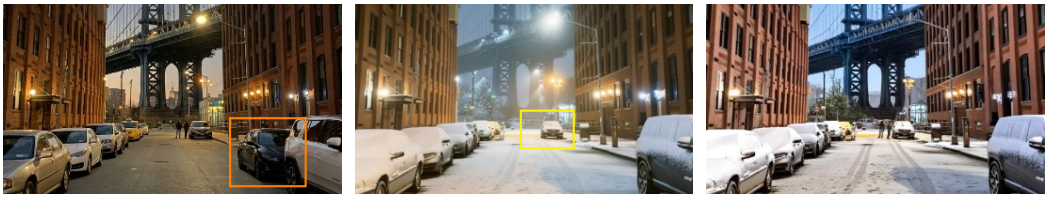
(a) Reference (b) Input (c) GPT-Image-2.0



(d) Nano Banana 2 (e) Qwen-Image-Edit (f) **RefRestore**



(a) Reference (b) Input (c) GPT-Image-2.0



(d) Nano Banana 2 (e) Qwen-Image-Edit (f) **RefRestore**

Figure 5: **Faithfulness failure of frontier MLLMs.** Colored boxes link regions in the input (b) to the method that corrupts it. RefRestore (f) preserves all observed content and matches the refrence (a), an unaligned clean photograph of the scene.

tent faithfulness rather than to sharpness or style, supporting their use as the evaluation protocol in Table 2.

A.4 Full Qualitative Results

Figures 1, 7, and 8 show RefRestore restorations on the full RefWeatherTest-20. Each restoration uses an unaligned clean reference as its appearance prior.

A.5 Full Qualitative Results

Figures 1, 7, and 8 show RefRestore results on the full RefWeatherTest-20 benchmark. Each restoration uses an unaligned clean image of the same scene as supplementary scene-specific information.

A.6 Discussion

Limitations. Two limitations are particularly relevant. First, RefRestore is evaluated primarily with references depicting the same scene and sharing useful visual content with the degraded observation. The proposed mechanisms reduce inappropriate transfer under this setting but do not guarantee rejection of an arbitrary or strongly mismatched reference. Because CARI retrieves according to feature similarity, a different-scene reference containing visually similar content, such

as facades or sky, may still influence the output in undesirable ways. Second, EGR uses channel variance as a lightweight relative evidence proxy. High-frequency degradations such as falling snow or rain streaks can increase feature variance and therefore be mistaken for informative content, reducing the intended suppression of those tokens and leaving residual degradation. More general estimation of observation reliability and explicit verification of transferred reference information remain open problems.

Future work. Two directions follow naturally from this work. First, RefUse and RefMisplace provide simple diagnostics for reference-associated content changes; extending such diagnostics toward spatially localized detection and verification of unsupported synthesized content would enable a more direct study of factuality in generative restoration. Second, the current framework controls reference influence through positional debiasing, evidence-based suppression, and optional semantic restrictions, but it does not verify that newly synthesized details are correct. Combining reference-assisted restoration with explicit consistency checks, multiple observations, or stronger scene verification mechanisms is therefore an important direction for future work.



(a) Reference (b) Input (c) 0.1522 / 0.5349 (d) 0.04815 / 0.2180



(a) Reference (b) Input (c) 0.1646 / 0.5382 (d) 0.1062 / 0.3129



(a) Reference (b) Input (c) 0.0463 / 0.3796 (d) 0.0153 / 0.1369



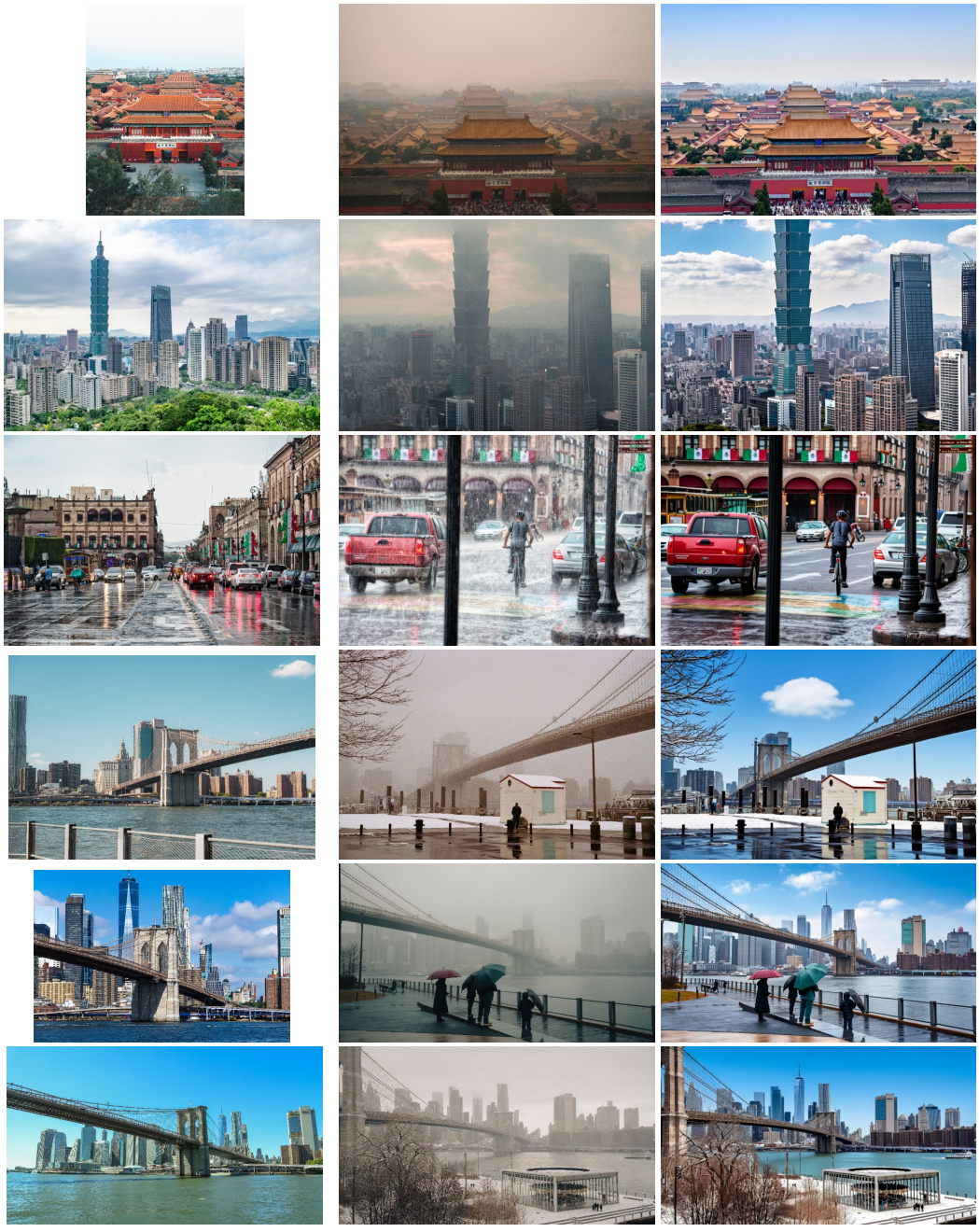
(a) Reference (b) Input (c) 0.2984 / 0.7046 (d) 0.0917 / 0.2929

Figure 6: **Sanity test of the faithfulness metrics.** Each scene is restored from the degraded input (b) with the unaligned reference (a) by naive reference injection (c) and by RefRestore (d), with RefUse / RefMisplace scores listed below each output.



(a) Reference (b) Input (c) **RefRestore** (d) Reference (e) Input (f) **RefRestore**

Figure 7: **Full results.** In each row we show the reference used, the degraded input, and our restored result.



(a) Reference (b) Input (c) **RefRestore**

Figure 8: **Full results.** In each row we show the reference used, the degraded input, and our restored result.