ROUTEFUSE: RELIABILITY-CONDITIONED ROUTING FOR MULTI-MODALITY IMAGE FUSION UNDER ADVERSE WEATHER

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modality image fusion integrates complementary sensor information, yet under adverse weather the need for infrared evidence varies spatially and across frequency bands because visible degradation is non-uniform and affects coarse structure and fine detail differently. Many existing methods rely on feature saliency or modality contribution, which do not directly measure local visible fidelity. We propose RouteFuse, a one-step diffusion-based framework that casts adverse-weather infrared–visible fusion as reliability-conditioned evidence allocation. RouteFuse represents infrared evidence as residuals relative to a primary multimodal representation and routes them separately in low- and high-frequency bands across multiple feature scales. Routing gates are conditioned on a learned visible reliability map and the cross-modal feature relation. A fusion-oriented diffusion learning scheme trains the predictor to estimate an explicit multimodal fusion target directly from sampled diffusion states, enabling one-step diffusion at inference. Mechanistic analyses show that predicted reliability aligns with local visible fidelity and that lower reliability is associated with stronger realized transfer of infrared residuals. Extensive experiments under rain, haze, and snow show that RouteFuse achieves state-of-the-art performance on all six primary fusion metrics with the lowest FLOPs and fastest inference among state-of-the-art competing methods.

1 INTRODUCTION

Multi-modality image fusion integrates complementary information captured by different sensors into a single image. Under adverse weather, however, infrared–visible image fusion (IVIF; Ma et al., 2019) must work with visible observations that rain, haze, and snow degrade through scattering, occlusion, contrast attenuation, and weather-induced artifacts (Li et al., 2026d;c; Liu et al., 2026a). This degradation is spatially non-uniform, leaving reliable scene content interspersed with corrupted observations. Because infrared observations provide thermal cues but not the appearance or color of visible images, fusion should preserve reliable visible content and draw on infrared evidence where visible observations become unreliable. The degradation also affects coarse structure and fine detail differently, so the need for infrared evidence varies both spatially and across frequency bands.

Recent IVIF methods strengthen multimodal interaction through feature aggregation, attention, long-range modeling, or semantic guidance (Xu et al., 2022; Ma et al., 2022; Yi et al., 2024), and adverse-weather methods further incorporate degradation modeling and restoration (Li et al., 2026b;c;d; Liu et al., 2026a). In these methods, cross-modal allocation is typically driven by feature saliency or by explicit estimates of modality contribution, for example from information measured on deep features (Xu et al., 2022) or from masks derived from the relation between fused and source images (Li et al., 2026b). Yet such cues describe how salient or informative each modality is, and neither property need coincide with the local fidelity of visible observations under adverse weather. As a result, allocation based on them may retain corrupted visible content or introduce infrared evidence where the visible observation is already reliable.

We therefore cast adverse-weather IVIF as reliability-conditioned evidence allocation and propose RouteFuse, a one-step diffusion-based framework that realizes this allocation through residual rout-

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.

ing. RouteFuse transfers infrared evidence represented as residuals relative to a primary multimodal representation, so that the evidence supplements rather than replaces this representation. Each residual is injected through a routing gate. The gate is conditioned on a learned reliability map that estimates local visible fidelity and on the cross-modal feature relation between the infrared features and the primary representation. Because coarse structure and fine detail call for different amounts of infrared evidence, low- and high-frequency residuals are routed separately at multiple feature scales. Residual routing is embedded in the encoder of a diffusion predictor. The predictor is trained to estimate an explicit multimodal fusion target directly from each sampled diffusion state, so it produces the fused output in one step at inference.

Under rain, haze, and snow, RouteFuse achieves the highest scores on all six primary fusion metrics, with the lowest FLOPs and fastest inference among state-of-the-art competing methods. Mechanistic analyses show that infrared transfer in RouteFuse follows local visible fidelity and that this allocation contributes to its fusion quality.

Our contributions are summarized as follows:

- We propose RouteFuse, a one-step diffusion-based framework that casts adverse-weather IVIF as reliability-conditioned evidence allocation.
- We design residual routing conditioned on visible reliability to selectively allocate complementary infrared evidence across frequency bands and feature scales, directing this evidence toward regions of low visible reliability.
- We develop a fusion-oriented diffusion learning scheme that trains the predictor to estimate an explicit multimodal fusion target directly from sampled diffusion states, enabling one-step diffusion at inference.

2 Related Work

2.1 Fusion under Adverse Weather

Degradation-aware fusion methods such as Text-IF (Yi et al., 2024) and OmniFuse (Zhang et al., 2025) mainly address low light, overexposure, and noise. ControlFusion (Tang et al., 2025) covers rain and haze, but only as two of its many degradation types. Methods designed for adverse weather mostly couple restoration with fusion. AWFusion (Li et al., 2026d) introduces the AWMM-100K benchmark, AMG-Fuse (Li et al., 2026b) uses mask-guided feature restoration and interaction, and CAWM-Mamba (Liu et al., 2026a) decouples compound weather degradations in the wavelet domain. AWM-Fuse (Li et al., 2026c) instead injects weather knowledge through text. UMCFuse (Li et al., 2025) separates degradation from scene content by transmission-guided layer decomposition. These methods focus on removing or describing the degradation. When they decompose frequencies, they do so to isolate degradations, not to decide how much infrared evidence each band should receive.

2.2 Adaptive Cross-Modal Allocation

PIAFusion (Tang et al., 2022) derives image-level weights from an estimated illumination probability to balance the modalities during training. TTD (Cao et al., 2024) assigns pixel-level weights at test time according to each modality's relative dominability in the fused image. Under degradation, TG-ECNet (Sun et al., 2025) selects experts with degradation-aware gates, and MdaIF (Li et al., 2026a) routes features to experts with language-driven semantic priors. Their allocation is driven by illumination, by each modality's contribution to the fused result, or by degradation cues. None of these directly measures the local fidelity of visible observations.

2.3 Diffusion-Based Fusion

DDFM (Zhao et al., 2023) and Dif-Fusion (Yue et al., 2023) establish diffusion-based fusion. Later methods introduce degradation-robust priors (Tang et al., 2024), text-modulated fusion (Zhang et al., 2024), joint fusion and super-resolution (Jie et al., 2025), projection-manifold regularization (Cao et al., 2025), and masked diffusion (Tang et al., 2026). Recent work further explores time-aware

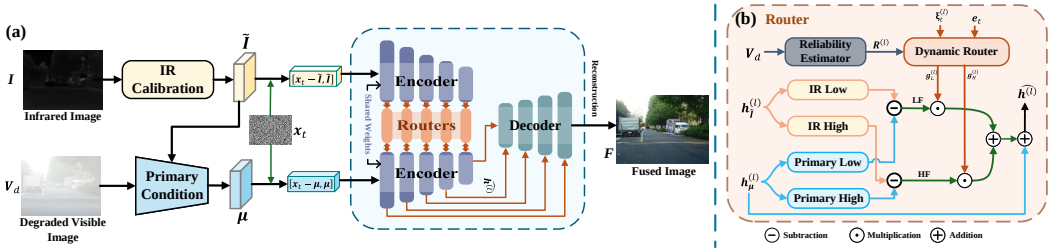


Figure 1: Overview of RouteFuse. (a) Overall framework, where routers inject infrared evidence into the primary representations at multiple feature scales. (b) Residual routing at feature scale ℓ, which uses the reliability map $\mathbf{R}^{(\ell)}$ to direct low- and high-frequency infrared residuals toward regions of low visible reliability. The projections $\mathcal{P}_b^{(\ell)}$ and $\mathcal{Q}^{(\ell)}$ are omitted for clarity.

cross-modal integration coupled with restoration (Zhang et al., 2026), diffusion-guided feature alignment (Wang et al., 2026), and direct regression of the fused image with observation correction during sampling (Shi et al., 2026). Most of these methods rely on iterative reverse sampling, which makes inference costly. One-step diffusion methods either straighten a pretrained rectified flow for image fusion (Wang et al., 2025) or distill a pretrained diffusion prior for video fusion (Liu et al., 2026b). Both build on pretrained generative models, and neither is designed for weather-degraded visible observations.

3 Method

3.1 Problem Formulation and Framework Overview

RouteFuse takes a degraded visible image $\mathbf{V}_d \in \mathbb{R}^{H\times W\times 3}$ and a spatially aligned infrared image $\mathbf{I} \in \mathbb{R}^{H\times W}$ as input. It produces a fused image $\mathbf{F}$ that suppresses weather degradation while preserving complementary information from both modalities. Fusion is performed on the luminance channel of $\mathbf{V}_d$, and the chrominance is restored by a separate branch. With a slight abuse of notation, $\mathbf{V}_d$ hereafter denotes the degraded visible luminance. Its clean counterpart $\mathbf{V}_{\mathrm{gt}}$ is available only during training.

As shown in Figure 1(a), RouteFuse realizes reliability-conditioned evidence allocation inside a diffusion predictor. The predictor estimates a multimodal fusion target from a sampled diffusion state $\mathbf{x}_t$. Its primary condition combines $\mathbf{V}_d$ with a calibrated infrared image $\tilde{\mathbf{I}}$, while a reliability map is predicted from $\mathbf{V}_d$ alone (Sec. 3.2). A shared encoder processes $\mathbf{x}_t$ twice, once under the primary condition and once under the infrared condition $\tilde{\mathbf{I}}$. Residual routing conditioned on visible reliability then injects infrared evidence from the infrared-conditioned representations into the primary representations at multiple feature scales (Sec. 3.3). Sec. 3.4 describes how the predictor is trained. At inference, RouteFuse produces the fused output from $\mathbf{V}_d$ and $\mathbf{I}$ in one step.

3.2 Visible Reliability and Conditioning

Adverse weather makes visible fidelity spatially non-uniform. RouteFuse therefore predicts a reliability map from the degraded visible luminance,

$$\mathbf{R} = \mathcal{R}_\phi(\mathbf{V}_d), \qquad \mathbf{R}^\star = \Psi(\mathcal{D}_{\mathrm{loc}}(\mathbf{V}_d, \mathbf{V}_{\mathrm{gt}})), \tag{1}$$

where $\mathcal{D}_{\mathrm{loc}}$ combines photometric and gradient discrepancies between the degraded and clean luminance, and $\Psi(\cdot)$ monotonically maps larger degradation to lower reliability. $\mathbf{R}^\star$ serves as the regression target of $\mathbf{R}$ during training. The resulting $\mathbf{R} \in [0,1]^{H\times W}$ estimates local visible fidelity rather than feature saliency or modality contribution, and is predicted from $\mathbf{V}_d$ alone at inference.

The primary condition combines the two modalities according to infrared saliency. A learnable global affine calibration maps the infrared image to $\tilde{\mathbf{I}}$, whose local intensity and gradient responses yield a soft infrared saliency map $\mathbf{M}$. The primary condition is

$$\boldsymbol{\mu} = (1-\mathbf{M}) \odot \mathbf{V}_d + \mathbf{M} \odot \tilde{\mathbf{I}}. \tag{2}$$

Visible reliability is not used in $\boldsymbol{\mu}$ and instead conditions residual routing (Sec. 3.3).

During training, a variance-preserving diffusion kernel (Ho et al., 2020; Song et al., 2021b) perturbs the multimodal fusion target $\mathbf{x}_0$ defined in Sec. 3.4,

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \tag{3}$$

where α_t and σ_t follow the noise schedule (Nichol & Dhariwal, 2021), $\boldsymbol{\epsilon}$ is standard Gaussian noise, and t is the timestep.

The shared encoder processes $\mathbf{x}_t$ once under the primary condition $\boldsymbol{\mu}$ and once under the infrared condition $\tilde{\mathbf{I}}$. With the timestep embedding $\mathbf{e}_t$, for $\mathbf{c} \in \{\boldsymbol{\mu}, \tilde{\mathbf{I}}\}$,

$$\left\{\mathbf{h}_{\mathbf{c},t}^{(\ell)}\right\}_{\ell=1}^{5} = \mathcal{E}_\theta\left([\mathbf{x}_t - \mathbf{c}, \mathbf{c}], \mathbf{e}_t\right). \tag{4}$$

The encoder receives both the state-to-condition discrepancy $\mathbf{x}_t - \mathbf{c}$ and the condition $\mathbf{c}$. Sharing $\mathbf{x}_t$, $\mathbf{e}_t$, and θ across the two passes makes the primary representations $\mathbf{h}_{\boldsymbol{\mu},t}^{(\ell)}$ and the infrared-conditioned representations $\mathbf{h}_{\tilde{\mathbf{I}},t}^{(\ell)}$ directly comparable, a property that residual routing relies on (Sec. 3.3).

3.3 Reliability-Conditioned Residual Routing

Residual routing transfers infrared evidence as residuals relative to the primary representation. Because the need for infrared evidence varies across spatial locations, frequency bands, and feature scales, the amount of transfer is decided separately for each. Routing is applied at four encoder skip scales and the bottleneck. Figure 1(b) illustrates residual routing at a single feature scale ℓ.

At scale ℓ and timestep t, both representations are decomposed with a fixed binomial low-pass operator $\mathcal{L}(\cdot)$. For $\mathbf{c} \in \{\boldsymbol{\mu}, \tilde{\mathbf{I}}\}$,

$$\mathbf{h}_{\mathbf{c},t}^{(\ell,\mathrm{L})} = \mathcal{L}\left(\mathbf{h}_{\mathbf{c},t}^{(\ell)}\right), \qquad \mathbf{h}_{\mathbf{c},t}^{(\ell,\mathrm{H})} = \mathbf{h}_{\mathbf{c},t}^{(\ell)} - \mathbf{h}_{\mathbf{c},t}^{(\ell,\mathrm{L})}. \tag{5}$$

The low-frequency component captures coarse variation, and the high-frequency component captures fine local responses, so coarse structure and fine detail can receive different amounts of infrared evidence.

For $b \in \{\mathrm{L}, \mathrm{H}\}$, the infrared residual is

$$\mathbf{r}_{b,t}^{(\ell)} = \mathcal{P}_b^{(\ell)}\left(\mathbf{h}_{\tilde{\mathbf{I}},t}^{(\ell,b)} - \mathbf{h}_{\boldsymbol{\mu},t}^{(\ell,b)}\right), \tag{6}$$

where $\mathcal{P}_b^{(\ell)}$ is a band-specific linear projection. The residual thus encodes how the infrared-conditioned representation departs from the primary representation within each band. It is neither a complete infrared feature nor a difference between the two source images.

Each band receives a spatial–channel routing gate,

$$[\mathbf{g}_{\mathrm{L},t}^{(\ell)}, \mathbf{g}_{\mathrm{H},t}^{(\ell)}] = \mathrm{sigmoid}\left(\mathcal{G}^{(\ell)}(\boldsymbol{\xi}_t^{(\ell)}, \mathbf{R}^{(\ell)}, \mathbf{e}_t)\right), \tag{7}$$

where $\mathbf{R}^{(\ell)}$ is the reliability map resized to scale ℓ. The cross-modal feature relation descriptor $\boldsymbol{\xi}_t^{(\ell)}$ is computed from the primary and infrared-conditioned representations and includes normalized responses, discrepancies, and band-wise statistics. The three inputs play complementary roles. $\mathbf{R}^{(\ell)}$ indicates where the visible observation can be trusted, $\boldsymbol{\xi}_t^{(\ell)}$ indicates how the two representations relate at each location, and $\mathbf{e}_t$ adapts the gates to the timestep.

The routed representation is

$$\hat{\mathbf{h}}_t^{(\ell)} = \mathbf{h}_{\boldsymbol{\mu},t}^{(\ell)} + \mathcal{Q}^{(\ell)}\left(\mathbf{g}_{\mathrm{L},t}^{(\ell)} \odot \mathbf{r}_{\mathrm{L},t}^{(\ell)} + \mathbf{g}_{\mathrm{H},t}^{(\ell)} \odot \mathbf{r}_{\mathrm{H},t}^{(\ell)}\right), \tag{8}$$

where $\mathcal{Q}^{(\ell)}$ maps the aggregated residual back to the space of the primary representation. The primary representation keeps an identity path, so infrared evidence enters only through the gated residuals. Because $\mathcal{Q}^{(\ell)}$ is affine, the injected term splits into its bias and one contribution per band.

With $\mathcal{Q}_0^{(\ell)}$ denoting $\mathcal{Q}^{(\ell)}$ without its bias, the contribution of band b is $\mathbf{u}_{b,t}^{(\ell)} = \mathcal{Q}_0^{(\ell)}\big(\mathbf{g}_{b,t}^{(\ell)} \odot \mathbf{r}_{b,t}^{(\ell)}\big)$. We measure the realized transfer of band b at location p as the channel-wise root mean square of this contribution,

$$\mathbf{T}_{b,t}^{(\ell)}(p) = \frac{1}{\sqrt{C}} \left\| \mathbf{u}_{b,t}^{(\ell)}(p) \right\|_2, \tag{9}$$

where $\mathbf{u}_{b,t}^{(\ell)}(p) \in \mathbb{R}^C$ collects the C channels of $\mathbf{u}_{b,t}^{(\ell)}$ at location p. Realized transfer thus depends jointly on the gate and on the residual available at each band and scale, rather than on the gate alone.

3.4 Fusion-Oriented Diffusion Learning

Residual routing determines how infrared evidence enters the predictor, while the learning objective specifies what the predictor should output. Fusion-oriented diffusion learning takes the desired fusion result as the learning target and constructs it explicitly as a multimodal fusion target.

The clean visible luminance $\mathbf{V}_{\mathrm{gt}}$ and the infrared image $\mathbf{I}_{\mathrm{gt}}$ are each decomposed into a low-frequency base and multi-scale band-pass components. The relative local texture energy of the two modalities determines a soft weight $\mathbf{A}_k$ at each scale $k \in \{1, \ldots, K\}$, and the target is

$$\mathbf{x}_0 = \mathrm{clip}_{[0,1]} \left[\max\Big(\mathbf{V}_{\mathrm{gt}}^{(K)}, \mathbf{I}_{\mathrm{gt}}^{(K)}\Big) + \sum_{k=1}^{K} \Big((1 - \mathbf{A}_k) \odot \mathbf{B}_V^{(k)} + \gamma_k \mathbf{A}_k \odot \mathbf{B}_I^{(k)} \Big) \right], \tag{10}$$

where $\mathbf{V}_{\mathrm{gt}}^{(K)}$ and $\mathbf{I}_{\mathrm{gt}}^{(K)}$ are the low-frequency bases, $\mathbf{B}_V^{(k)}$ and $\mathbf{B}_I^{(k)}$ are the band-pass components, and γ_k controls the infrared contribution at scale k. The target thus takes the element-wise maximum of the low-frequency bases and selects fine details from both modalities according to local texture.

Using the v-prediction parameterization (Salimans & Ho, 2022), we define $\mathbf{v}_t^\star = \alpha_t \boldsymbol{\epsilon} - \sigma_t \mathbf{x}_0$ and reconstruct $\hat{\mathbf{x}}_0 = \alpha_t \mathbf{x}_t - \sigma_t \hat{\mathbf{v}}_t$. The predicted velocity is supervised toward $\mathbf{v}_t^\star$, and $\hat{\mathbf{x}}_0$ is compared with $\mathbf{x}_0$ through an L1 intensity term. This image-space term pulls the prediction from every sampled diffusion state toward the same fusion target. Because the predictor is trained to reconstruct $\mathbf{x}_0$ directly from each state, it produces the fused output in one step, without iterative reverse sampling.

On adverse-weather samples, two further terms compare $\hat{\mathbf{x}}_0$ with $\mathbf{x}_0$. $\mathcal{L}_{\mathrm{str}}$ supervises signed first-order variations across multiple spatial scales, and $\mathcal{L}_{\mathrm{tex}}$ supervises the high-frequency subbands of a two-level Haar decomposition. Clean samples use modified forms of both terms. Because gradients of image-space losses with respect to $\hat{\mathbf{v}}_t$ scale with σ_t, both terms are weighted according to the sampled training timestep. The reliability map is regressed toward $\mathbf{R}^\star$ with a Smooth-L1 loss ($\mathcal{L}_{\mathrm{rel}}$). The full objective is

$$\mathcal{L} = \|\hat{\mathbf{v}}_t - \mathbf{v}_t^\star\|_2^2 + \lambda_{\mathrm{int}} \|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|_1 + \lambda_{\mathrm{str}} \mathcal{L}_{\mathrm{str}} + \lambda_{\mathrm{tex}} \mathcal{L}_{\mathrm{tex}} + \lambda_{\mathrm{rel}} \mathcal{L}_{\mathrm{rel}} + \mathcal{L}_{\mathrm{aux}}, \tag{11}$$

where $\mathcal{L}_{\mathrm{aux}}$ collects three auxiliary terms. They supervise directional Haar-detail predictions from auxiliary decoder heads, preserve the edges of the clean visible and infrared images, and keep local brightness consistent with $\mathbf{x}_0$. The Appendix gives all definitions and loss weights, including the forms used for clean samples.

At inference, $\mathbf{x}_{299}$ is drawn from a standard Gaussian and the predictor is applied once at $t = 299$. The fused output is $\hat{\mathbf{x}}_0$ clipped to $[0, 1]$, which corresponds to a one-step DDIM update with $\eta = 0$ (Song et al., 2021a).

4 Experiments

4.1 Experimental Setup

RouteFuse is trained on registered infrared–visible pairs from AWMM-100K (Li et al., 2026d) with the Lion optimizer (Chen et al., 2023), 256×256 random crops, and a batch size of 4. For adverse-weather evaluation, 50 training-disjoint scene groups are reserved for each of rain, haze, and snow, covering light, moderate, and heavy degradation. Scores for each weather type are averaged over its three severity levels, and overall scores are averaged over all nine weather–severity conditions. All methods are evaluated at their original spatial resolution. Fusion quality is measured

by Q_{MI} (Hossny et al., 2008), Q_G and $Q_{\mathrm{AB/F}}$ (Xydeas & Petrovic, 2000a;b), VIF (Han et al., 2013), SSIM (Wang et al., 2004), and SCD (Aslantas & Bendes, 2015).

Among the baselines, CAWM-Mamba (Liu et al., 2026a) and AMG-Fuse (Li et al., 2026b) are trained on AWMM-100K, the same dataset as RouteFuse, and directly process the degraded infrared–visible pairs. UMCFuse (Li et al., 2025) is parameter-free and also processes the degraded pairs directly. For Text-IF, Text-DiFuse, OmniFuse, ReCoFuse, DRMF, and Mask-DiFuser (Yi et al., 2024; Zhang et al., 2024; 2025; 2026; Tang et al., 2024; 2026), adverse-weather inputs are first restored with WeatherRemover (WR) (Qu et al., 2026) and then fused. Author-released checkpoints are used for all baselines. Further implementation details and results under clean conditions on the M3FD fusion test set are provided in the Appendix.

4.2 Main Results

Adverse weather. As Table 1(a) shows, RouteFuse achieves the highest scores on all six primary fusion metrics. Its closest competitors are CAWM-Mamba and AMG-Fuse, which are trained on the same AWMM-100K data as RouteFuse. The gains over them are largest on SCD, VIF, and SSIM. Table 1(b) breaks the results down by weather type, and the complete six-metric results are given in the Appendix. RouteFuse achieves the highest SSIM and SCD under every weather type and the highest Q_{MI} under rain and snow.

Table 1: Comparison under adverse weather. (a) Overall results. (b) Results of RouteFuse and the three best-ranked competitors in (a) under each weather type, with complete results in the Appendix. Bold and underline mark the best and second-best results.

(a) Overall adverse-weather comparison

Method	$Q_{\mathrm{MI}}\uparrow$	$Q_G\uparrow$	$Q_{\mathrm{AB/F}}\uparrow$	VIF↑	SSIM↑	SCD↑
WR + Mask-DiFuser	0.2906	0.2555	0.3521	0.6481	0.4576	1.3453
WR + Text-IF	0.3333	0.2162	0.3121	0.4360	0.4602	0.9387
WR + DRMF	0.3313	0.1842	0.2813	0.5430	0.4856	0.9316
WR + Text-DiFuse	0.2590	0.1863	0.2797	0.4821	0.3783	0.8663
WR + ReCoFuse	0.3007	0.1899	0.2618	0.4353	0.3949	0.8381
WR + OmniFuse	0.3172	0.1708	0.2441	0.3891	0.3808	0.8247
UMCFuse	0.2370	0.1934	0.2218	0.2142	0.4170	0.6937
AMG-Fuse	0.3930	0.3743	0.5387	<u>0.7274</u>	0.6452	1.3561
CAWM-Mamba	<u>0.4352</u>	<u>0.3977</u>	<u>0.5636</u>	0.7252	<u>0.6585</u>	<u>1.3849</u>
RouteFuse	**0.4529**	**0.4162**	**0.5760**	**0.7952**	**0.7005**	**1.6516**

(b) Weather-wise comparison

Weather	Method	$Q_{\mathrm{MI}}\uparrow$	SSIM↑	SCD↑
Rain	WR + Mask-DiFuser	0.2831	0.4553	<u>1.5012</u>
	AMG-Fuse	0.4090	0.6550	1.4354
	CAWM-Mamba	<u>0.4378</u>	<u>0.6739</u>	1.4101
	RouteFuse	**0.4673**	**0.7098**	**1.6965**
Haze	WR + Mask-DiFuser	0.2866	0.4360	0.9716
	AMG-Fuse	0.3764	<u>0.6629</u>	1.0826
	CAWM-Mamba	**0.4622**	0.6581	<u>1.1880</u>
	RouteFuse	<u>0.4348</u>	**0.6685**	**1.5359**
Snow	WR + Mask-DiFuser	0.3020	0.4815	<u>1.5632</u>
	AMG-Fuse	0.3935	0.6178	1.5503
	CAWM-Mamba	<u>0.4056</u>	<u>0.6435</u>	1.5566
	RouteFuse	**0.4566**	**0.7232**	**1.7223**

Figure 2 shows representative results for each weather type. RouteFuse suppresses weather artifacts while keeping pedestrian contours and scene structure intact. Several competing methods leave residual degradation or introduce local intensity and texture distortions.

Downstream detection. Following CAWM-Mamba (Liu et al., 2026a), we train YOLO11x on the original training set of AWMM-100K and keep it fixed. The detector is then applied without fine-tuning to the fused images of all methods on the annotated adverse-weather images of AWMM-100K. RouteFuse achieves the highest precision, recall, mAP@0.5, and mAP@0.5:0.95 (Table 2). The margin is largest on recall, so fewer objects are missed in weather-degraded scenes.

Efficiency. Because RouteFuse needs only one step at inference, it is also efficient. Table 3 reports costs at 640×480 resolution on an NVIDIA RTX 5880 Ada GPU, where WR-based pipelines include WR (24.3M parameters, 0.74T FLOPs, 0.15 s per image). RouteFuse requires the fewest FLOPs and runs fastest, at about half the runtime of WR + Text-IF. UMCFuse is parameter-free and runs on a CPU, so it is excluded from the parameter ranking.

4.3 Mechanistic Analysis of Reliability-Conditioned Residual Routing

Residual routing is designed to transfer more infrared evidence where visible observations are less reliable. We test whether predicted reliability reflects local visible fidelity, whether lower reliability is associated with stronger realized transfer, and whether the resulting allocation matters for fusion quality.



Figure 2: Qualitative comparison under rain, haze, and snow. Zoomed regions highlight weather suppression, contrast recovery, and structural preservation. WR-based methods first restore the degraded visible image with WR before fusion.

Table 2: Object detection on fused images under adverse weather. Values are in %. Bold and underline mark the best and second-best results.

Method	Precision↑	Recall↑	mAP@0.5↑	mAP@0.5:0.95↑
WR + Mask-DiFuser	86.19	70.98	79.82	54.39
WR + Text-IF	86.98	69.08	78.48	53.61
WR + DRMF	81.68	65.52	73.64	49.43
WR + Text-DiFuse	84.52	68.54	76.51	51.91
WR + ReCoFuse	83.94	68.51	76.70	51.70
WR + OmniFuse	83.22	62.03	71.02	47.72
UMCFuse	83.06	63.90	74.18	49.99
AMG-Fuse	<u>87.54</u>	70.38	79.12	53.87
CAWM-Mamba	86.96	<u>75.16</u>	<u>83.55</u>	<u>58.01</u>
RouteFuse	**87.84**	**78.65**	**85.64**	**60.17**

We measure local visible fidelity by comparing $\mathbf{V}_d$ with $\mathbf{V}_{\mathrm{gt}}$ through a gradient-based measure and local SSIM. Predicted reliability correlates positively with both. The mean within-image Spearman correlation is $\rho = 0.203$ for the gradient-based measure and $\rho = 0.344$ for local SSIM (Figure 3(a)).

We then relate reliability to the realized transfer $\mathbf{T}_b$ in Eq. (9), computed on the four skip scales and both frequency bands. Across all eight routing branches, lower reliability is associated with stronger realized transfer. The mean within-image Spearman correlations range from -0.480 to -0.148, and all 95% cluster-bootstrap confidence intervals exclude zero (Figure 3(b)). The association is

Table 3: Efficiency comparison of all methods. Bold and underline mark the best and second-best results.

Method	Params (M)	FLOPs (T)	Time (s/image)
WR + Mask-DiFuser	196.648	94.392	4.7724
WR + Text-IF	239.438	3.979	<u>0.4063</u>
WR + DRMF	195.305	77.916	3.3057
WR + Text-DiFuse	143.777	575.087	15.1448
WR + ReCoFuse	179.795	4.561	4.7553
WR + OmniFuse	197.661	4.212	4.0831
UMCFuse	0	N/A	6.1683
AMG-Fuse	**60.073**	<u>3.300</u>	0.8731
CAWM-Mamba	<u>120.904</u>	5.543	0.6649
RouteFuse	213.000	**1.960**	**0.1998**

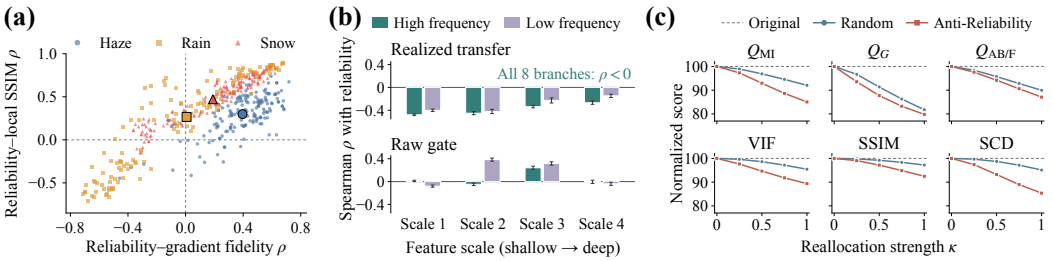


Figure 3: Mechanistic analysis of residual routing. (a) Per-image correlations between predicted reliability and the two fidelity measures, with weather-wise means shown as large markers. (b) Mean within-image Spearman correlations of reliability with realized transfer and with raw gates for the eight routing branches. (c) Energy-matched Anti-Reliability and Random reallocations. Scores are normalized to the original model (100%).

generally stronger at shallow scales and is strongest at the shallowest high-frequency branch. Raw gate values, in contrast, vary in sign and magnitude across branches. This is expected, because the injected amount depends on the residual as well as on the gate. Figure 4 shows the same tendency spatially, with distinct patterns for low- and high-frequency transfer.

Correlation alone does not show that the allocation matters. We therefore redistribute routed-residual strength across locations while keeping the energy of each intervened branch unchanged. Anti-Reliability scales down injection by a factor of $1-\kappa$ at the 20% of locations with the lowest reliability and scales it up at the 20% with the highest reliability. A matched Random control applies the same rule to disjoint random regions of equal size. At every tested nonzero strength, Anti-Reliability yields lower scores than Random on all six metrics (Figure 3(c)). Since both interventions inject the same energy and differ only in where it goes, this gap indicates that reliability-conditioned spatial allocation matters for fusion quality.

4.4 Ablation Studies

Removing any of the three routing components (A1–A3) lowers all six metrics (Table 4). Reliability conditioning (A1) tells the gates where visible observations are unreliable. Frequency decomposition (A2) lets coarse structure and fine detail receive different amounts of infrared evidence, while the timestep embedding (A3) adapts the gates to the timestep. The largest drop on every metric comes from removing residual routing altogether (A4). When the gates are replaced by trained static ones (A5), they no longer adapt to the input, and all six metrics decline.

Using a pixel-wise maximum of $\mathbf{V}_{\mathrm{gt}}$ and $\mathbf{I}_{\mathrm{gt}}$ as the target (A6) discards the selection of fine details according to local texture, and all six metrics fall. Uniform structure–texture weights (A7), which ignore the sampled training timestep, lead to the same outcome. A8 keeps the same network architecture, including residual routing, but predicts the fused output deterministically, without a sampled

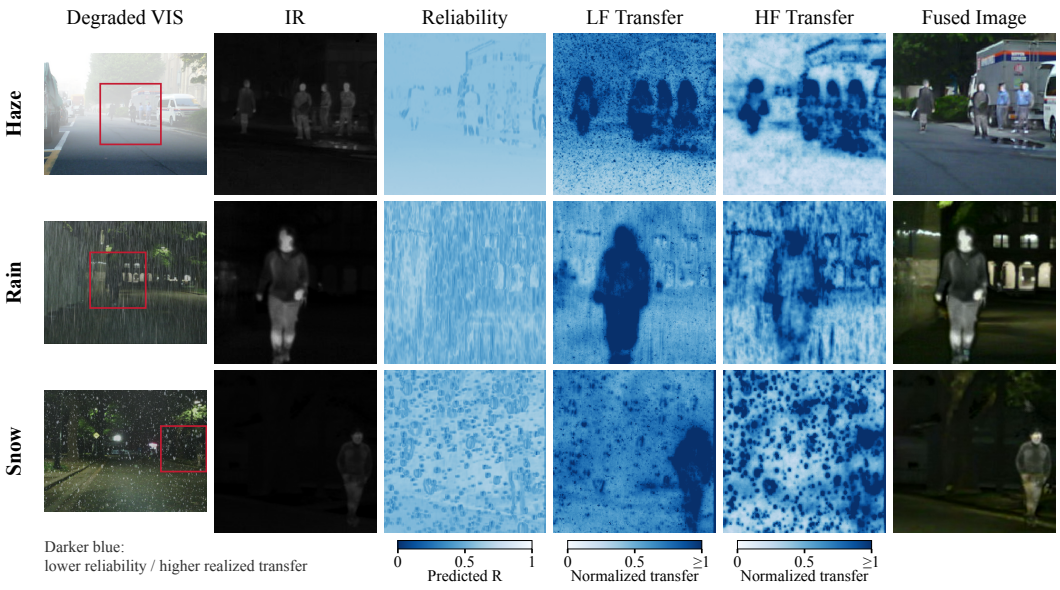


Figure 4: Spatial maps of predicted reliability and realized infrared transfer. Low-frequency (LF) and high-frequency (HF) maps aggregate the four feature scales with fixed dataset-level branch normalization.

Table 4: Ablation results under adverse weather. Bold marks the best results.

Variant	$Q_{\mathrm{MI}}\uparrow$	$Q_G\uparrow$	$Q_{\mathrm{AB/F}}\uparrow$	VIF↑	SSIM↑	SCD↑
A1: w/o Reliability Conditioning	0.4217	0.3871	0.5460	0.7458	0.6903	1.5370
A2: w/o Frequency Decomposition	0.4179	0.3839	0.5433	0.7361	0.6913	1.5289
A3: w/o Gate Timestep Conditioning	0.4228	0.3937	0.5511	0.7534	0.6862	1.5441
A4: w/o Residual Routing	0.3769	0.3231	0.4944	0.6844	0.6792	1.4221
A5: Trained Static Gates	0.4460	0.4085	0.5672	0.7817	0.6997	1.6213
A6: Pixel-wise Max Target	0.4272	0.4001	0.5541	0.7511	0.6835	1.5546
A7: Uniform Structure–Texture Weighting	0.4218	0.3775	0.5361	0.7206	0.6936	1.5418
A8: Deterministic Encoder–Decoder	0.3926	0.4142	0.5554	0.7639	0.6863	1.5613
Full Model	**0.4529**	**0.4162**	**0.5760**	**0.7952**	**0.7005**	**1.6516**

diffusion state. It falls below the full model on all six metrics, so under the same architecture the diffusion formulation outperforms deterministic prediction.

5 Conclusion

We propose RouteFuse, a one-step diffusion-based framework that casts adverse-weather infrared–visible image fusion as reliability-conditioned evidence allocation. RouteFuse represents infrared evidence as residuals relative to a primary multimodal representation and routes them separately in low- and high-frequency bands across multiple feature scales. Routing is conditioned on a learned visible reliability map and the cross-modal feature relation, directing infrared evidence toward regions of low reliability. This ties the use of infrared evidence to local visible fidelity rather than to feature saliency or modality contribution. We further develop a fusion-oriented diffusion learning scheme that trains RouteFuse to estimate an explicit multimodal fusion target directly from sampled diffusion states, enabling one-step diffusion at inference. Experiments under rain, haze, and snow show that RouteFuse achieves state-of-the-art fusion performance with the lowest FLOPs and the fastest inference among state-of-the-art competing methods.

AI USE STATEMENT

We used large language models to aid and polish the writing of this paper, and to assist with retrieving and discovering related work. All content was reviewed and verified by the authors, who take full responsibility for this paper.

REPRODUCIBILITY STATEMENT

The architecture and training objective of RouteFuse are described in Sec. 3.1–3.4, and the evaluation protocol is given in Sec. 4.1. Appendix A and Appendix B provide the implementation details, loss definitions, and loss weights. Anonymized inference code and pretrained weights are provided as supplementary material.

REFERENCES

Veysel Aslantas and Emre Bendes. A new image quality metric for image fusion: The sum of the correlations of differences. *AEU - International Journal of Electronics and Communications*, 69 (12):1890–1896, 2015. doi: 10.1016/j.aeue.2015.09.004.

Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Test-time dynamic image fusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 2080–2105, 2024. doi: 10.52202/079017-0067.

Lei Cao, Hao Zhang, Chunyu Li, and Jiayi Ma. Projection-manifold regularized latent diffusion for robust general image fusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 93568–93599, 2025. doi: 10.52202/085713-2814.

Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision – ECCV 2022*, volume 13667 of *Lecture Notes in Computer Science*, pp. 17–33, Cham, 2022. Springer Nature Switzerland. doi: 10.1007/978-3-031-20071-7_2.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms. In *Advances in Neural Information Processing Systems*, volume 36, pp. 49205–49233, 2023. doi: 10.52202/075280-2140.

Yu Han, Yunze Cai, Yin Cao, and Xiaoming Xu. A new image fusion performance metric based on visual information fidelity. *Information Fusion*, 14(2):127–135, 2013. doi: 10.1016/j.inffus.2011.08.002.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

M. Hossny, S. Nahavandi, and D. Creighton. Comments on 'information measure for performance of image fusion'. *Electronics Letters*, 44(18):1066–1067, 2008. doi: 10.1049/el:20081754.

Yuchan Jie, Yushen Xu, Xiaosong Li, Fuqiang Zhou, Jianming Lv, and Huafeng Li. FS-Diff: Semantic guidance and clarity-aware simultaneous multimodal image fusion and super-resolution. *Information Fusion*, 121:103146, 2025. doi: 10.1016/j.inffus.2025.103146.

Jing Li, Yifan Wang, Jiafeng Yan, Renlong Zhang, and Bin Yang. MdaIF: Robust one-stop multi-degradation-aware image fusion with language-driven semantics. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(8):6226–6234, 2026a. doi: 10.1609/aaai.v40i8.37548.

Xilai Li, Xiaosong Li, Tianshu Tan, Huafeng Li, and Tao Ye. UMCFuse: A unified multiple complex scenes infrared and visible image fusion framework. *IEEE Transactions on Image Processing*, 34: 6231–6245, 2025. doi: 10.1109/TIP.2025.3607623.

Xilai Li, Xiaosong Li, Haishu Tan, Tao Ye, Huafeng Li, and Hongbin Wang. Multi-modality image fusion under adverse weather: Mask-guided feature restoration and interaction. In *Computer Vision – ECCV 2026*, volume 17027 of *Lecture Notes in Computer Science*, pp. 112–130, Cham, 2026b. Springer Nature Switzerland. doi: 10.1007/978-3-032-37602-2_7.

Xilai Li, Huichun Liu, Xiaosong Li, Tao Ye, Zhenyu Kuang, and Huafeng Li. AWM-Fuse: Multi-modality image fusion for adverse weather via global and local text perception. *IEEE Transactions on Image Processing*, 35:5151–5164, 2026c. doi: 10.1109/TIP.2026.3690324.

Xilai Li, Wuyang Liu, Xiaosong Li, Fuqiang Zhou, Huafeng Li, and Feiping Nie. All-weather multi-modality image fusion: Unified framework and 100k benchmark. *Information Fusion*, 131: 104130, 2026d. doi: 10.1016/j.inffus.2026.104130.

Huichun Liu, Xiaosong Li, Zhuangfan Huang, Tao Ye, Yang Liu, and Haishu Tan. CAWM-Mamba: A unified model for infrared-visible image fusion and compound adverse weather restoration. *Expert Systems with Applications*, 314:131665, 2026a. doi: 10.1016/j.eswa.2026.131665.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5802–5811, 2022. doi: 10.1109/CVPR52688.2022.00571.

Jinyuan Liu, Ludan Sun, Tengyu Ma, Chunyan Yang, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Streaming diffusion model for fast infrared and visible video fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14305–14314, 2026b.

Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019. doi: 10.1016/j.inffus.2018.02.004.

Jiayi Ma, Linfeng Tang, Fan Fan, Jun Huang, Xiaoguang Mei, and Yong Ma. SwinFusion: Cross-domain long-range learning for general image fusion via Swin Transformer. *IEEE/CAA Journal of Automatica Sinica*, 9(7):1200–1217, 2022. doi: 10.1109/JAS.2022.105686.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

Weikai Qu, Sijun Liang, Cheng Pan, Zikuan Yang, Guanchi Zhou, Xianjun Fu, Bo Liu, Changmiao Wang, and Ahmed Elazab. WeatherRemover: All-in-one adverse weather removal with multiscale feature map compression. *IEEE Transactions on Artificial Intelligence*, 7(5):2980–2994, 2026. doi: 10.1109/TAI.2025.3633206.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Yu Shi, Yu Liu, Zhong-Cheng Wu, Juan Cheng, Huafeng Li, and Xun Chen. Degradation-robust fusion: An efficient degradation-aware diffusion framework for multimodal image fusion in arbitrary degradation scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33848–33858, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yiming Sun, Xin Li, Pengfei Zhu, Qinghua Hu, Dongwei Ren, Huiying Xu, and Xinzhong Zhu. Task-gated multi-expert collaboration network for degraded multi-modal image fusion. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57571–57586. PMLR, 2025. URL https://proceedings.mlr.press/v267/sun25k.html.

Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. PIAFusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83–84:79–92, 2022. doi: 10.1016/j.inffus.2022.03.007.

Linfeng Tang, Yuxin Deng, Xunpeng Yi, Qinglong Yan, Yixuan Yuan, and Jiayi Ma. DRMF: Degradation-robust multi-modal image fusion via composable diffusion prior. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8546–8555, 2024. doi: 10.1145/3664647.3681064.

Linfeng Tang, Yeda Wang, Zhanchuan Cai, Junjun Jiang, and Jiayi Ma. ControlFusion: A controllable image fusion network with language-vision degradation prompts. In *Advances in Neural Information Processing Systems*, volume 38, pp. 131142–131165, 2025. doi: 10.52202/085713-3946.

Linfeng Tang, Chunyu Li, and Jiayi Ma. Mask-DiFuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48 (1):591–608, 2026. doi: 10.1109/TPAMI.2025.3609323.

Yuequn Wang, Zhengwei Li, Leqiang Yang, Wei Xu, Tianyu Lv, and Jie Liu. DiffFuseNet: Infrared-visible image fusion via diffusion-guided feature alignment and feature consistency alignment. *Pattern Recognition*, 180:114259, 2026. doi: 10.1016/j.patcog.2026.114259.

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.

Zirui Wang, Jiayi Zhang, Tianwei Guan, Yuhan Zhou, Xingyuan Li, Minjing Dong, and Jinyuan Liu. Efficient rectified flow for image fusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 26019–26040, 2025. doi: 10.52202/085713-0774. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/212b143b5a5d6b88feb0fb1441b9756e-Abstract-Conference.html.

Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 502–518, 2022. doi: 10.1109/TPAMI.2020.3012548.

Costas S. Xydeas and Vladimir Petrovic. Objective image fusion performance measure. *Electronics Letters*, 36(4):308–309, 2000a. doi: 10.1049/el:20000267.

Costas S. Xydeas and Vladimir S. Petrovic. Objective pixel-level image fusion performance measure. In *Sensor Fusion: Architectures, Algorithms, and Applications IV*, volume 4051, pp. 89–98. SPIE, 2000b. doi: 10.1117/12.381668.

Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-IF: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27026–27035, 2024. doi: 10.1109/CVPR52733.2024.02552.

Jun Yue, Leyuan Fang, Shaobo Xia, Yue Deng, and Jiayi Ma. Dif-Fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models. *IEEE Transactions on Image Processing*, 32:5705–5720, 2023. doi: 10.1109/TIP.2023.3322046.

Hao Zhang, Lei Cao, and Jiayi Ma. Text-DiFuse: An interactive multi-modal image fusion framework based on text-modulated diffusion model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 39552–39572, 2024. doi: 10.52202/079017-1250.

Hao Zhang, Lei Cao, Xuhui Zuo, Zhenfeng Shao, and Jiayi Ma. OmniFuse: Composite degradation-robust image fusion with language-driven semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7577–7595, 2025. doi: 10.1109/TPAMI.2025.3568433.

Hao Zhang, Shuhan Yang, Linfeng Tang, Xunpeng Yi, and Jiayi Ma. ReCoFuse: Ultra-robust image fusion via restorative multi-modal diffusion reciprocal coupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33838–33847, 2026.

Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. DDFM: Denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8082–8093, 2023. doi: 10.1109/ICCV51070.2023.00742.

Appendix

A Implementation Details

Architecture. The predictor is a time-conditioned U-Net-style encoder–decoder built on NAFNet (Chen et al., 2022), with four skip scales and a bottleneck, and a router is placed at each of these five feature scales. In each router, $\mathcal{Q}^{(\ell)}$ consists of a bias-free depthwise 3×3 convolution followed by a 1×1 convolution with bias, with no nonlinearity in between. The reliability estimator R_ϕ is a lightweight convolutional head that outputs a sigmoid map, and the chrominance branch is also based on NAFNet. Two auxiliary detail heads on the decoder are used only during training.

Training and inference. RouteFuse is trained with the Lion optimizer for 300K iterations on one NVIDIA RTX 5880 Ada GPU, using 256×256 random crops and a batch size of 4. Module-wise learning rates between 1×10^{-6} and 1×10^{-5} are used with cosine annealing. The diffusion process uses $T = 300$ timesteps with the cosine noise schedule (Nichol & Dhariwal, 2021). At inference, $\mathbf{x}_{299}$ is drawn from a standard Gaussian, the predictor is applied once at $t = 299$, and $\hat{\mathbf{x}}_0$ is clipped to $[0, 1]$.

B Training Objective

Throughout this section, $\rho(\cdot)$ denotes the Charbonnier penalty, $\mathcal{D}_s$ denotes downsampling by a factor s, ∂_h and ∂_v denote horizontal and vertical differences, and $(\omega_1, \omega_2, \omega_4) = (1, 0.5, 0.25)$ are scale weights with sum 1.75.

Fusion target. The decomposition used for the fusion target $\mathbf{x}_0$ is a full-resolution Gaussian-residual decomposition with $K = 3$ band-pass components and one low-frequency base. At each scale, $\mathbf{A}_k$ is determined by the local texture energy of $\mathbf{B}_I^{(k)}$ relative to that of $\mathbf{B}_V^{(k)}$.

Reliability target. Let $\mathbf{d}_{\text{int}}$ and $\mathbf{d}_{\text{grad}}$ be the intensity and gradient discrepancies between $\mathbf{V}_d$ and $\mathbf{V}_{\text{gt}}$. Each is normalized and smoothed with a 9×9 Gaussian kernel with a standard deviation of 2. The regression target is

$$\mathbf{R}^\star = \exp(-2\mathbf{d}), \qquad \mathbf{d} = 0.7\,\mathbf{d}_{\text{int}} + 0.3\,\mathbf{d}_{\text{grad}}, \tag{12}$$

and $\mathcal{L}_{\text{rel}}$ is the Smooth-L1 loss between $\mathbf{R}$ and $\mathbf{R}^\star$.

Structure and texture terms. On adverse-weather samples, $\mathcal{L}_{\text{str}}$ compares signed first-order variations of $\hat{\mathbf{x}}_0$ and $\mathbf{x}_0$ at three scales,

$$\mathcal{L}_{\text{str}} = \frac{1}{1.75} \sum_{s \in \{1,2,4\}} \frac{\omega_s}{2} \sum_{d \in \{h,v\}} \text{mean} \left| \partial_d \mathcal{D}_s(\hat{\mathbf{x}}_0) - \partial_d \mathcal{D}_s(\mathbf{x}_0) \right|, \tag{13}$$

and $\mathcal{L}_{\text{tex}}$ applies ρ to the differences between the high-frequency subbands of two-level Haar decompositions of $\hat{\mathbf{x}}_0$ and $\mathbf{x}_0$.

On clean samples, $\mathcal{L}_{\text{str}}$ compares $\hat{\mathbf{x}}_0$ with both $\mathbf{V}_{\text{gt}}$ and $\mathbf{I}_{\text{gt}}$ through SSIM,

$$\mathcal{L}_{\text{str}}^{\text{clean}} = 0.75\,\mathcal{L}_{\text{MS}} + 0.75\,\lambda_G\,\mathcal{L}_G. \tag{14}$$

Here $\mathcal{L}_{\text{MS}}$ averages $1 - \text{SSIM}$ over the two sources and the scales $s \in \{1, 2, 4\}$ with an 11×11 uniform window, and $\mathcal{L}_G$ averages $1 - \text{SSIM}$ over the two sources at full resolution with an 11×11 Gaussian window with a standard deviation of 1.5. The factor $\lambda_G = 0.68$ matches the gradient norm of $\mathcal{L}_G$ with respect to $\hat{\mathbf{x}}_0$ to that of $\mathcal{L}_{\text{MS}}$ and is computed once on clean training crops before training. $\mathcal{L}_{\text{tex}}$ on clean samples penalizes only gradients of $\hat{\mathbf{x}}_0$ that exceed those of both sources,

$$\mathcal{L}_{\text{tex}}^{\text{clean}} = \text{mean}\left[\text{ReLU}\Big(|\nabla \hat{\mathbf{x}}_0| - \max\left(|\nabla \mathbf{V}_{\text{gt}}|, |\nabla \mathbf{I}_{\text{gt}}|\right)\Big)\right], \tag{15}$$

where $|\nabla \cdot|$ is the Sobel gradient magnitude.

Timestep weighting. Since $\hat{\mathbf{x}}_0 = \alpha_t \mathbf{x}_t - \sigma_t \hat{\mathbf{v}}_t$, the gradient of an image-space loss with respect to $\hat{\mathbf{v}}_t$ is scaled by σ_t. $\mathcal{L}_{\text{str}}$, $\mathcal{L}_{\text{tex}}$, and $\mathcal{L}_{\text{edge}}$ are therefore multiplied by

$$w(t) = \frac{1}{Z} \min\Big(w_{\max}, \frac{c}{\sigma_t}\Big), \qquad Z = \frac{1}{T} \sum_{\tau=0}^{T-1} \min\Big(w_{\max}, \frac{c}{\sigma_\tau}\Big), \tag{16}$$

where $c = 0.5$ and $w_{\max} = 8$. The normalization keeps the mean weight over all timesteps at one. The ablation A7 in Table 4 replaces $w(t)$ with 1 for $\mathcal{L}_{\text{str}}$ and $\mathcal{L}_{\text{tex}}$ and keeps it for $\mathcal{L}_{\text{edge}}$, with all loss weights unchanged.

Auxiliary terms. $\mathcal{L}_{\text{aux}} = 0.10\,\mathcal{L}_{\text{det}} + 0.40\,\mathcal{L}_{\text{edge}} + 0.05\,\mathcal{L}_{\text{bri}}$. In $\mathcal{L}_{\text{det}}$, the intermediate and final detail heads predict the LH, HL, and HH Haar subbands of $\mathbf{x}_0$, and their Charbonnier losses are combined with weights 0.5 and 1.0 and normalized by their sum. $\mathcal{L}_{\text{edge}}$ compares signed gradients of $\hat{\mathbf{x}}_0$ with those of $\mathbf{V}_{\text{gt}}$ and $\mathbf{I}_{\text{gt}}$ at scales $s \in \{1, 2, 4\}$. At each pixel, the two comparisons are weighted by a softmax with temperature 10 over the local gradient magnitudes of the two sources, so the stronger source edge dominates,

$$\begin{aligned}\mathcal{L}_{\text{edge}} = \frac{1}{1.75} \sum_s \frac{\omega_s}{2} \sum_{d \in \{h,v\}} \text{mean} \Big[\mathbf{w}_V \, \rho\big(\partial_d \mathcal{D}_s(\hat{\mathbf{x}}_0) - \partial_d \mathcal{D}_s(\mathbf{V}_{\text{gt}})\big) \\ + \mathbf{w}_I \, \rho\big(\partial_d \mathcal{D}_s(\hat{\mathbf{x}}_0) - \partial_d \mathcal{D}_s(\mathbf{I}_{\text{gt}})\big) \Big].\end{aligned} \tag{17}$$

$\mathcal{L}_{\text{bri}}$ is the L1 distance between 16×16 average-pooled versions of $\hat{\mathbf{x}}_0$ and $\mathbf{x}_0$.

Loss weights. Table 5 summarizes all loss weights.

Table 5: Loss weights.

$\|\hat{\mathbf{v}}_t - \mathbf{v}_t^\star\|_2^2$	$\|\hat{\mathbf{x}}_0 - \mathbf{x}_0\|_1$	$\mathcal{L}_{\text{str}}$	$\mathcal{L}_{\text{tex}}$	$\mathcal{L}_{\text{rel}}$	$\mathcal{L}_{\text{det}}$	$\mathcal{L}_{\text{edge}}$	$\mathcal{L}_{\text{bri}}$
1.00	0.15	0.10	0.08	0.05	0.10	0.40	0.05

C Additional Results

Results under each weather type. Table 6 reports all six metrics under each weather type, averaged over the three severity levels.

Results under clean conditions. Table 7 reports results on clean inputs from the M3FD fusion test set (Liu et al., 2022). RouteFuse uses the same checkpoint as in the adverse-weather experiments without fine-tuning, and its training split does not overlap the M3FD test scenes. The baselines fuse the clean inputs directly. RouteFuse achieves the highest SSIM and the second-highest Q_{MI}, Q_G, and $Q_{\text{AB/F}}$.

D Energy-Matched Reallocation

The interventions act on the band contribution $\mathbf{u}_b^{(\ell)} = \mathcal{Q}_0^{(\ell)}(\mathbf{g}_b^{(\ell)} \odot \mathbf{r}_b^{(\ell)})$ of each of the eight routing branches, separately for each image. Let Ω_{low} and Ω_{high} be the 20% of locations with the lowest and highest reliability, and let E_{low} and E_{high} be the sums of $\|\mathbf{u}_b^{(\ell)}(p)\|_2^2$ over them. Anti-Reliability multiplies $\mathbf{u}_b^{(\ell)}$ by $1 - \kappa$ on Ω_{low} and by

$$\beta = \sqrt{1 + \big[1 - (1-\kappa)^2\big] \frac{E_{\text{low}}}{E_{\text{high}}}} \tag{18}$$

on Ω_{high}, and leaves the remaining locations unchanged. This keeps the total squared norm $\sum_p \|\mathbf{u}_b^{(\ell)}(p)\|_2^2$ of the branch unchanged. The Random control applies the same rule to two disjoint random regions of the same size. Model weights, inputs, initial noise, and residual-vector locations are fixed for both interventions.

Table 6: Results under each weather type. Bold and underline mark the best and second-best results.

Weather	Method	$Q_{\mathrm{MI}}\uparrow$	$Q_G\uparrow$	$Q_{\mathrm{AB/F}}\uparrow$	VIF↑	SSIM↑	SCD↑
Rain	WR + Mask-DiFuser	0.2831	0.2480	0.4304	**0.8529**	0.4553	<u>1.5012</u>
	WR + Text-IF	0.3331	0.2565	0.4436	0.6686	0.5353	1.0840
	WR + DRMF	0.3225	0.2227	0.3845	0.7998	0.4312	0.9806
	WR + Text-DiFuse	0.2552	0.2328	0.4058	0.7562	0.3720	1.0328
	WR + ReCoFuse	0.3108	0.2061	0.3363	0.5311	0.3739	0.8797
	WR + OmniFuse	0.3029	0.2103	0.3571	0.6014	0.3929	0.8951
	UMCFuse	0.2340	0.1960	0.2846	0.2878	0.4838	0.7074
	AMG-Fuse	0.4090	0.3281	0.5078	0.7875	0.6550	1.4354
	CAWM-Mamba	<u>0.4378</u>	<u>0.3340</u>	<u>0.5170</u>	0.7687	<u>0.6739</u>	1.4101
	RouteFuse	**0.4673**	**0.3828**	**0.5636**	<u>0.8503</u>	**0.7098**	**1.6965**
Haze	WR + Mask-DiFuser	0.2866	0.2512	0.2181	0.2550	0.4360	0.9716
	WR + Text-IF	0.3037	0.1661	0.1222	0.1021	0.3894	0.6386
	WR + DRMF	0.3088	0.1229	0.1032	0.1212	0.4598	0.6523
	WR + Text-DiFuse	0.2133	0.1216	0.0969	0.0937	0.3916	0.4668
	WR + ReCoFuse	0.2729	0.1553	0.1518	0.2040	0.4226	0.6890
	WR + OmniFuse	0.2995	0.1100	0.0794	0.0845	0.3789	0.6068
	UMCFuse	0.2888	0.2256	0.1414	0.1135	0.4872	0.7767
	AMG-Fuse	0.3764	0.4844	0.6255	0.6587	<u>0.6629</u>	1.0826
	CAWM-Mamba	**0.4622**	**0.5493**	**0.6763**	<u>0.6618</u>	0.6581	<u>1.1880</u>
	RouteFuse	<u>0.4348</u>	<u>0.5004</u>	<u>0.6342</u>	**0.7095**	**0.6685**	**1.5359**
Snow	WR + Mask-DiFuser	0.3020	0.2672	0.4077	**0.8362**	0.4815	<u>1.5632</u>
	WR + Text-IF	0.3631	0.2260	0.3706	0.5372	0.4559	1.0934
	WR + DRMF	0.3625	0.2070	0.3564	0.7079	0.5658	1.1618
	WR + Text-DiFuse	0.3086	0.2044	0.3363	0.5964	0.3714	1.0993
	WR + ReCoFuse	0.3184	0.2084	0.2972	0.5709	0.3883	0.9456
	WR + OmniFuse	0.3493	0.1920	0.2959	0.4814	0.3707	0.9722
	UMCFuse	0.1883	0.1584	0.2394	0.2413	0.2801	0.5971
	AMG-Fuse	0.3935	<u>0.3104</u>	0.4828	0.7359	0.6178	1.5503
	CAWM-Mamba	<u>0.4056</u>	0.3099	<u>0.4974</u>	0.7451	<u>0.6435</u>	1.5566
	RouteFuse	**0.4566**	**0.3655**	**0.5300**	<u>0.8259</u>	**0.7232**	**1.7223**

Table 7: Results under clean conditions on the M3FD fusion test set. Bold and underline mark the best and second-best results.

Method	$Q_{\mathrm{MI}}\uparrow$	$Q_G\uparrow$	$Q_{\mathrm{AB/F}}\uparrow$	VIF↑	SSIM↑	SCD↑
Mask-DiFuser	0.3733	0.3507	0.4316	**1.2018**	0.5814	**1.8219**
Text-IF	0.5341	**0.5665**	**0.6760**	0.4998	0.7090	<u>1.5188</u>
DRMF	**0.6653**	0.3422	0.5210	0.3400	0.6834	0.9934
Text-DiFuse	0.3870	0.3867	0.5312	0.4642	0.6812	1.3324
ReCoFuse	0.4249	0.3071	0.4518	<u>0.6407</u>	0.6601	1.3497
OmniFuse	0.5048	0.2920	0.4345	0.3593	0.6980	1.2859
UMCFuse	0.4542	0.5272	0.5281	0.4565	<u>0.7138</u>	1.5094
AMG-Fuse	0.3253	0.4551	0.5818	0.5020	0.7083	1.5037
CAWM-Mamba	0.4194	0.4765	0.5773	0.5381	0.6864	1.3930
RouteFuse	<u>0.5816</u>	<u>0.5286</u>	<u>0.6435</u>	0.4087	**0.7190**	1.3828