Interactive All-in-One Image Restoration and Fusion

Bing Cao[1] , **Qiang Zhang**[1] , **Xingxin Xu**[2] and **Pengfei Zhu**[1,3,4,5*]
[1]School of Artificial Intelligence, Tianjin University
[2]School of New Media and Communication, Tianjin University
[3]Low-Altitude Intelligence Laboratory, Xiong'an National Innovation Center
[4]Xiong'an Guochuang Lantian Technology Co., Ltd.
[5]National Key Laboratory of Science and Technology on Automatic Target Recognition, National University of Defense Technology
{caobing,zhangqiang_,xuxingxin,zhupengfei}@tju.edu.cn

Abstract

Supervised infrared-visible image fusion (IVIF) often overfits limited training distributions, creating a critical generalization gap under open-world degradations (rain, haze, low light, noise, blur). To address this issue, we propose AIR-Fusion, a parameter-efficient adaptation of a frozen, restoration-capable latent diffusion backbone for degraded IVIF without full fine-tuning, transferring restoration priors for robust fusion. A Cross-Modal Bridging Adapter (CMBA) aligns infrared cues and textual instructions with the frozen diffusion conditioning space and injects them into multi-scale denoising features to steer instruction-guided restoration-aware fusion. In addition, a Trajectory-Constrained Rectifier (TCR) regularizes stochastic sampling via a pixel-latent closed loop, rectifying intermediate predictions with source-referenced structures and re-encoding them to stabilize the denoising trajectory and recover fine details suppressed by latent compression. Experiments across multiple datasets and degradation settings show consistent improvements in restoration quality and fusion fidelity, with strong generalization under complex and compounded degradations.

1 Introduction

Multi-modal image fusion integrates complementary cues from different sensors into a unified representation [Li and Wu, 2024; Cao *et al.*, 2024]. A prime example is Infrared-Visible Image Fusion (IVIF) [Zhang and Demiris, 2023], where infrared sensors capture thermal radiation to highlight salient targets [Ibarra-Castanedo *et al.*, 2004], while visible sensors provide rich textural and chromatic details. However, the modality gap makes it challenging to balance target prominence with fine-grained scene detail. Ideally, fusion should preserve thermal saliency and visible texture without artifacts, benefiting downstream perception such as object detection and surveillance [Takumi *et al.*, 2017].

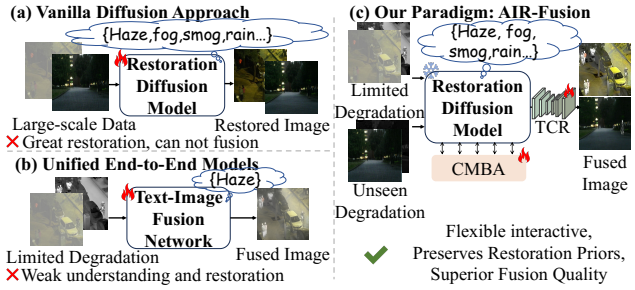

Figure 1: Comparative analysis of fusion paradigms. **(a)** Pre-trained models possess robust priors but lack multi-modal fusion capacity. **(b)** Unified methods trained from scratch often lack strong restoration priors and robustness under complex degradations. **(c)** Ours adapts a frozen backbone: CMBA elicits restoration-aware fusion, while TCR ensures structural fidelity via trajectory constraints.

Despite impressive progress, practical deployment faces a critical hurdle: supervised IVIF models are often trained on limited fusion datasets and can overfit the training distribution, resulting in a pronounced generalization gap under open-world degradations. In real scenarios, inputs are inevitably corrupted by complex degradations (e.g., rain, haze, noise, and blur) [Bulat *et al.*, 2018; Tang *et al.*, 2023]. Conventional fusion pipelines [Fang *et al.*, 2021; Zhao *et al.*, 2020] typically treat inputs as clean signals, inadvertently propagating and even amplifying degradations in the fused results. Such corruption distorts modality-specific structures and salient cues, leading to inconsistent textures and weakened target boundaries. Consequently, downstream perception may degrade in safety-critical scenarios. Thus, robust restoration-aware fusion is urgently needed to restore signal fidelity while integrating complementary information.

To address this, recent unified frameworks like Text-IF [Yi *et al.*, 2024b] attempt to couple restoration with fusion. However, most rely on training task-specific networks from scratch on limited fusion datasets. Compared with large-scale generative corpora, these datasets lack diversity, causing scratch-trained models to struggle with open-world degradations and varying imaging conditions. As illustrated in Fig. 1, pre-trained generative models possess robust priors

*Corresponding author.

for restoration but lack inherent multi-modal fusion capability, whereas unified methods trained from scratch lack strong visual priors for restoration. Bridging this gap requires leveraging the restoration priors of foundation models for generalizable multimodal fusion under complex degradations.

Motivated by these priors, we propose AIR-Fusion, a framework that adapts a frozen restoration-capable LDM backbone for robust fusion under degradations. Instead of training a fusion network from scratch, we reformulate degraded IVIF as aligning multi-modal conditions with this frozen generative substrate, so that restoration priors learned from large-scale data can generalize beyond the limited fusion training distribution. We term this reformulation generative capability elicitation. Yet fusion imposes stricter source-fidelity requirements than ordinary restoration: the model should remove degradations without hallucinating structures or weakening thermal targets. However, directly repurposing such a backbone for high-fidelity fusion is non-trivial due to three key gaps: *(i)* the backbone is trained for single-modal conditional generation and thus lacks a native multi-modal conditioning interface to jointly condition on degraded visible and infrared inputs; *(ii)* full adaptation risks degrading the pre-trained restoration priors that underpin open-world robustness; and *(iii)* stochastic sampling together with VAE latent compression can suppress high-frequency structures and induce structural drift, conflicting with the source fidelity requirements of fusion.

To overcome these hurdles, we employ a parameter-efficient fine-tuning (PEFT) strategy. We freeze this backbone and introduce a lightweight Cross-Modal Bridging Adapter (CMBA). CMBA maps infrared cues and text instructions into the frozen conditioning space and injects them into multi-scale denoising features, enabling instruction-guided restoration-aware fusion without expensive full-model training. Furthermore, to counteract the loss of high-frequency details caused by VAE-based latent compression in LDMs [Chen *et al.*, 2023] and regularize the stochastic denoising trajectory, we propose a Trajectory-Constrained Rectifier (TCR). TCR explicitly references infrared and visible raw source inputs for modality-aware rectification in a pixel-latent closed loop, constraining each denoising step toward source-consistent structures while recovering fine-grained details. The main contributions of this work are summarized as follows:

- We propose AIR-Fusion, an interactive all-in-one image restoration and fusion framework that performs generative capability elicitation for degraded IVIF by aligning multi-modal conditions with a frozen generative manifold to preserve restoration priors while eliciting fusion capability under complex degradations.
- We propose CMBA for cross-modal conditioning and TCR for trajectory regularization via a pixel-latent closed loop, jointly improving restoration-aware fusion and source fidelity.
- Extensive experiments on multiple datasets and real-world degradations show consistent gains in restoration quality, fusion fidelity, and cross-dataset robustness.

2 Related Work

Multi-modality Image Fusion. Deep learning has substantially advanced image fusion beyond traditional optimization-based pipelines [Shi *et al.*, 2005; Pradnya and Sachin, 2013], primarily due to its powerful representation learning capabilities. Early data-driven methods were dominated by CNN and auto-encoder architectures [Liu *et al.*, 2017; Zhang *et al.*, 2020; Jian *et al.*, 2020], followed by adversarial learning [Zhou *et al.*, 2023; Ma *et al.*, 2020] and Transformer-based mechanisms (e.g., SwinFusion [Ma *et al.*, 2022]) to enhance perceptual fidelity and capture long-range dependencies. However, a critical limitation of these networks is their implicit assumption of clean, high-quality inputs. Under degradations (e.g., rain, haze, noise, low-light and blur), such assumptions often break, leading to degraded fused outputs and weakened saliency/texture cues. To address degradations, recent studies have shifted towards unified restoration-and-fusion pipelines. Text-IF [Yi *et al.*, 2024b] pioneered the integration of textual prompts to guide interactive fusion, while ControlFusion [Tang *et al.*, 2025] and MMAIF [Cao *et al.*, 2025] further incorporated degradation-aware prompts. Despite their progress, these approaches predominantly adhere to a training-from-scratch paradigm; constrained by limited fusion data, they often exhibit a generalization gap under complex, open-world degradations.

Diffusion Models and Foundation Model Adaptation. Denoising Diffusion Probabilistic Models (DDPMs) [Ho *et al.*, 2020; Nichol and Dhariwal, 2021] and Latent Diffusion Models (LDMs) [Rombach *et al.*, 2022] provide strong generative priors for solving inverse problems. (i) *Generative Priors in Restoration.* Leveraging large-scale pre-training, LDMs encode rich statistics of natural images. Recent works like AutoDIR [Jiang *et al.*, 2024] and Spire [Qi *et al.*, 2024] show that robust blind restoration can be elicited from frozen generative priors. Nevertheless, these frameworks are tailored for single-modality restoration and lack mechanisms to integrate and align complementary information from multi-modal sensors when the backbone is kept frozen. (ii) *Diffusion-based Fusion.* Diffusion architectures have also been applied to image fusion, such as DDFM [Zhao *et al.*, 2023b], Diff-IF [Yi *et al.*, 2024a], and Text-DiFuse [Zhang *et al.*, 2024]. However, most existing approaches treat diffusion models as trainable backbones and adapt extensively on limited fusion data, which can underutilize a frozen restoration-capable manifold and weaken robustness under severe degradations. In contrast, leveraging a frozen restoration manifold for fusion calls for dedicated multi-modal condition alignment rather than full-model training. (iii) *Foundation Model Adaptation.* Parameter-efficient fine-tuning (PEFT) techniques like ControlNet [Zhang *et al.*, 2023] and T2I-Adapter [Mou *et al.*, 2024] inject conditional signals via lightweight modules while freezing heavy backbones. While effective for controlled generation, robust degraded fusion further requires preserving restoration priors, aligning multi-modal conditions, and enforcing source fidelity during sampling. AIR-Fusion is designed to bridge these gaps via generative capability elicitation with condition alignment (CMBA) and trajectory-level fidelity constraints (TCR).

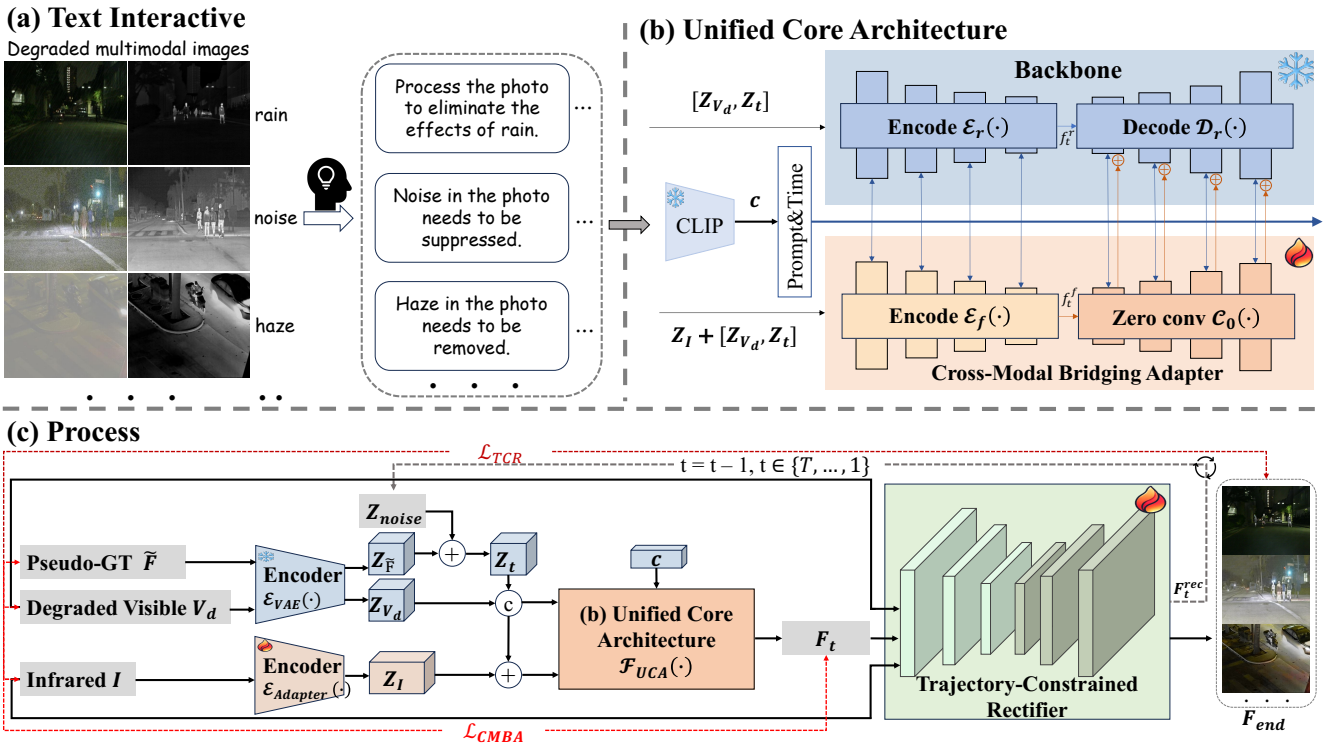

Figure 2: The overall framework of AIR-Fusion. Adopting a foundation model adaptation paradigm, the CMBA adapter elicits multi-modal fusion capabilities from a frozen restoration-capable LDM backbone via interactive latent diffusion. Subsequently, the TCR module iteratively rectifies the generated results to ensure high-fidelity structural preservation.

3 Methodology

The architecture of AIR-Fusion, illustrated in Fig. 2, is founded on the principle of generative capability elicitation. Diverging from traditional paradigms that attempt to reconstruct fusion mappings from limited data, our framework adopts a transfer learning strategy. Concretely, in this work we instantiate this substrate with AutoDIR [Jiang *et al.*, 2024], a restoration-capable latent diffusion backbone rather than a vanilla LDM, and keep it frozen to preserve its pre-trained restoration priors. Upon this foundation, we introduce and exclusively optimize lightweight modules to align multi-modal conditions and constrain the denoising trajectory for robust fusion under degradations.

The overall pipeline comprises two synergistic stages. First, the Condition Alignment stage employs a Cross-Modal Bridging Adapter (CMBA) to align multi-modal conditions with the frozen backbone, steering the iterative denoising process via textual prompts. Second, the Trajectory-Constrained Rectification stage utilizes a Trajectory-Constrained Rectifier (TCR) module to enforce source-referenced fidelity, mitigating the spectral bias introduced by VAE latent compression and significantly enhancing fine-grained visual quality.

3.1 CMBA Adapter for Fusion

We formulate the problem using Visible-Infrared Fusion as a representative case study. Let $I_{\text{raw}} \in \mathbb{R}^{H\times W}$ denote the raw infrared image and $V \in \mathbb{R}^{H\times W\times 3}$ denote the clean visible image. In practice, for compatibility with the backbone, we replicate the infrared modality channel-wise to obtain $I \in \mathbb{R}^{H\times W\times 3}$ ($I = \text{Replicate}(I_{\text{raw}})$), and we use I to denote this 3-channel representation throughout the paper. In real-world scenarios, the visible input is often corrupted, yielding a degraded observation V_d, where the degradation type $d \in \{\text{low-light, rain, haze, blur, noise}\}$.

We employ AutoDIR [Jiang *et al.*, 2024] as the frozen restoration-capable latent-diffusion backbone. The degraded image V_d is projected into a latent representation $Z_{V_d} \in \mathbb{R}^{h\times w\times c}$ via the pre-trained VAE encoder:

$$Z_{V_d} = \mathcal{E}_{VAE}(V_d). \tag{1}$$

To inject complementary infrared information without disrupting the pre-trained generative manifold, we introduce the Cross-Modal Bridging Adapter (CMBA). CMBA is inspired by ControlNet-style conditional branches [Zhang *et al.*, 2023]: it is a trainable conditional branch whose condition input is the infrared modality. Specifically, CMBA encodes the infrared input I into a latent condition Z_I that lies in the same latent space as Z_{V_d} and is spatially aligned with it, while using twice the channel capacity:

$$Z_I = \mathcal{E}_{Adapter}(I), \tag{2}$$

where $Z_I \in \mathbb{R}^{h\times w\times 2c}$. During denoising, CMBA produces multi-scale residual features that are injected back into the frozen backbone through zero-initialized 1×1 convolutions and are added to the corresponding decoder blocks.

Training Strategy: Manifold Alignment. Image fusion is inherently unsupervised due to the lack of ground-truth fused labels. Following [Yi *et al.*, 2024a], we construct a curated repository of reference priors, denoted as $\mathcal{F}_g$, comprising high-quality fused images generated by representative algorithms. At training time, we sample a proxy fused prior $\tilde{F} \sim \mathcal{F}_g$ and encode it as $Z_{\tilde{F}} = \mathcal{E}_{VAE}(\tilde{F})$. We then sample a timestep $t \sim \text{Uniform}(1, T)$ and add standard Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ to obtain the noisy latent:

$$Z_t = \sqrt{\bar{\alpha}_t}\, Z_{\tilde{F}} + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \tag{3}$$

where $Z_t \in \mathbb{R}^{h\times w\times c}$. Meanwhile, user text prompts (e.g., "A clear photo, no haze") are encoded into embeddings c via a frozen CLIP text encoder to provide semantic guidance.

The feature extraction process follows two parallel pathways. Let $[\cdot;\cdot]$ denote channel-wise concatenation and $\oplus$ denote element-wise addition. The Frozen Restoration Path employs the frozen backbone U-Net encoder $\mathcal{E}_r$ to encode the concatenation of the current noisy latent Z_t and the degraded-visible latent Z_{V_d}, producing restoration-prior features f_t^r:

$$f_t^r = \mathcal{E}_r\big([Z_t; Z_{V_d}],\, t,\, c\big). \tag{4}$$

In parallel, the Trainable Fusion Path uses the CMBA encoder $\mathcal{E}_f$, which further incorporates the infrared latent Z_I by adding it to the concatenated visible-conditioned latent, yielding fusion-oriented features f_t^f:

$$f_t^f = \mathcal{E}_f\big([Z_t; Z_{V_d}] \oplus Z_I,\, t,\, c\big). \tag{5}$$

Within the decoder of the frozen backbone U-Net ($\mathcal{D}_r$), the fusion features are projected by a zero-initialized convolution (denoted as $\mathcal{C}_0(\cdot)$) and injected at multiple scales via element-wise addition, yielding a steered representation for noise prediction:

$$\hat{\epsilon}_t = \mathcal{D}_r\big(f_t^r \oplus \mathcal{C}_0(f_t^f),\, t,\, c\big). \tag{6}$$

Synergistic Inference Process. During inference, the Unified Core Architecture $\mathcal{F}_{UCA}$ (the denoising core that couples the frozen backbone with the CMBA) starts from pure Gaussian noise $Z_T \sim \mathcal{N}(0, \mathbf{I})$ and performs iterative denoising, unlike training where Z_t is constructed by noising a proxy-prior latent. At each timestep, $\mathcal{F}_{UCA}$ predicts $\hat{\epsilon}_t$ to update the latent according to the diffusion scheduler.

This design fosters a mutual reinforcement mechanism: the fusion cues from CMBA provide modality-complementary guidance, while the frozen restoration-capable LDM backbone contributes strong degradation removal priors. The process yields a preliminary fused result that is semantically consistent but may exhibit high-frequency loss due to latent compression, motivating the subsequent trajectory-constrained rectification stage.

3.2 Trajectory-Constrained Rectification

A known limitation of LDMs [Chen *et al.*, 2023; Huang *et al.*, 2023] is that the VAE-induced spatial compression may suppress high-frequency textures, which is particularly harmful for image fusion where source fidelity is critical. To mitigate this issue, we propose a Trajectory-Constrained Rectifier (TCR), denoted as $\mathcal{R}$, which explicitly references raw source inputs to enforce a structure-preserving constraint in pixel space. Architecturally, TCR adapts the refinement design from the NAFNet block [Chen *et al.*, 2022]; crucially, we extend the input to enable multi-modal awareness.

We decode the clean-latent estimate $\hat{Z}_{0|t}$ into a coarse fused image:

$$\hat{Z}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}\left(Z_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\epsilon}_t\right). \tag{7}$$

$$F_t = \mathcal{D}_{VAE}(\hat{Z}_{0|t}), \tag{8}$$

and feed it to TCR together with the degraded visible image V_d and infrared image I via channel-wise concatenation:

$$F_t^{rec} = \mathcal{R}\big([F_t; V_d; I]\big), \tag{9}$$

Since F_t already contains the main clean fusion content, TCR uses V_d mainly as a structural reference under final fused-image supervision rather than copying its degradation artifacts.

To maximally leverage TCR, we embed it directly into the sampling loop, forming a pixel-latent closed loop that regularizes the denoising trajectory. At each timestep t, we first estimate the clean latent $\hat{Z}_{0|t}$ and decode it to pixel space. TCR then rectifies this estimate, which is subsequently re-encoded to guide the next sampling step:

$$\hat{Z}_{0|t}^{rec} = \mathcal{E}_{VAE}(F_t^{rec}) \tag{10}$$

$$Z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\hat{Z}_{0|t}^{rec} + \sqrt{1-\bar{\alpha}_{t-1}}\,\hat{\epsilon}_t. \tag{11}$$

By iteratively enforcing this constraint, AIR-Fusion anchors the generative trajectory to the physical structures of the source inputs, yielding high-fidelity fusion results under severe degradations.

3.3 Loss Function

Our training process is divided into two sequential stages, and our loss function is designed accordingly.

In the first stage, we train CMBA and frozen backbone. The primary objective is to learn the denoising process fundamental to the diffusion model. This is formulated as a noise prediction loss, $\mathcal{L}_{noise}$:

$$\mathcal{L}_{noise} = \mathbb{E}_{\tilde{F}\sim\mathcal{F}_g,\, t,\, \epsilon}\left[\|\epsilon - \hat{\epsilon}_t\|_2^2\right]. \tag{12}$$

To guide the model towards better fusion results and enhance visual quality, we introduce a set of auxiliary fusion losses: a Structural Similarity loss ($\mathcal{L}_{ssim}$), an intensity loss ($\mathcal{L}_{int}$), a gradient loss ($\mathcal{L}_{grad}$), and a color loss ($\mathcal{L}_{color}$). The total loss for this stage is a weighted sum of the primary and auxiliary objectives:

$$\begin{aligned}\mathcal{L}_{CMBA} = \sigma_{noise}\mathcal{L}_{noise} + \sigma_{ssim}\mathcal{L}_{ssim} + \sigma_{int}\mathcal{L}_{int} \\ +\sigma_{grad}\mathcal{L}_{grad} + \sigma_{color}\mathcal{L}_{color}.\end{aligned} \tag{13}$$

In the second stage, we freeze CMBA and optimize TCR. The training objective, $\mathcal{L}_{TCR}$, leverages the same set of auxiliary fusion losses. Crucially, these losses are now computed on the final output image to directly refine its detail and visual quality.

Methods	Blur				Haze				Low-light				Rain				Noise			
	PSNR	SSIM	VIF	LPIPS	PSNR	SSIM	VIF	LPIPS	PSNR	SSIM	VIF	LPIPS	PSNR	SSIM	VIF	LPIPS	PSNR	SSIM	VIF	LPIPS
U2Fusion*	27.98	0.58	0.44	0.43	27.66	0.51	0.62	0.43	28.01	0.54	0.50	0.46	28.33	0.61	0.48	0.40	28.23	0.72	0.56	0.41
TarDAL*	28.15	0.54	0.45	0.47	27.99	0.52	0.53	0.47	27.92	0.51	0.46	0.48	27.79	0.48	0.50	0.51	27.83	0.66	0.50	0.47
SwinFusion*	28.39	0.56	0.53	0.43	28.46	0.41	0.28	0.56	28.00	0.47	0.35	0.53	28.62	0.57	0.47	0.44	28.82	0.67	0.49	0.47
CDDFuse*	28.54	0.58	0.62	0.42	27.90	0.49	0.39	0.47	28.00	0.48	0.41	0.50	28.62	0.59	0.49	0.43	28.04	0.61	0.53	0.49
DDFM*	28.11	0.59	0.52	0.43	28.47	0.60	0.73	0.39	28.15	0.55	0.52	0.47	28.62	0.64	0.57	0.39	27.85	0.68	0.58	0.43
EMMA*	28.08	0.56	0.51	0.43	27.85	0.48	0.31	0.48	27.86	0.47	0.35	0.51	28.71	0.60	0.50	0.42	27.87	0.61	0.53	0.49
Text-IF	29.34	0.60	0.77	0.41	29.16	0.59	0.78	0.40	28.64	0.55	0.57	0.49	29.18	0.60	0.78	0.40	28.78	0.70	0.81	0.39
ControlFusion	29.46	0.57	0.65	0.43	29.28	0.57	0.67	0.42	28.97	0.56	0.61	0.47	29.33	0.63	0.70	0.41	28.81	0.62	0.56	0.37
AIR-Fusion(Ours)	29.56	0.61	0.77	0.40	29.53	0.61	0.73	0.38	29.21	0.57	0.69	0.44	30.54	0.67	0.79	0.38	29.08	0.66	0.58	0.40

Table 1: Quantitative comparison under degraded conditions. Columns correspond to blur (LLVIP), haze (LLVIP), low-light (LLVIP), rain (MSRS), and noise (RoadScene). Red: best; Blue: second best.

4 Experiments

Datasets. Our model was trained exclusively on the LLVIP dataset [Jia *et al.*, 2021]. To comprehensively evaluate its effectiveness and generalization capability, we utilized the MSRS [Tang *et al.*, 2022] and RoadScene [Xu *et al.*, 2020] datasets as unseen test sets. For the training data, we leveraged both the inherent degradations found in LLVIP's visible images (e.g., low light, blur) and further applied a suite of five artificially simulated degradations: low light, rain, haze, blur, and noise. We construct a template-based prompt set (hundreds of variants) covering the five degradation types to enable instruction-guided interaction. In addition, we use AWMM-100k [Li *et al.*, 2024] for qualitative evaluation of real-world generalization under adverse weather.

Implementation Details. Our method utilizes the pre-trained restoration-capable AutoDIR latent-diffusion backbone [Jiang *et al.*, 2024] as its frozen backbone. During training, we optimize only the parameters of the CMBA and the TCR. Parameter updates are performed using the Adam optimizer with a learning rate set to 1e-5. All experiments were conducted on an NVIDIA RTX A6000 GPU. With 100 denoising steps, AIR-Fusion has only 0.361B trainable parameters and takes 11.397s for inference, requiring far fewer trainable parameters than full AutoDIR tuning while running much faster than Diff-IF and DDFM.

Evaluation Metrics. To quantitatively assess the quality of the fused images, we employed a variety of evaluation metrics: SD [Ma *et al.*, 2019], EN [Ma *et al.*, 2019], VIF [Han *et al.*, 2013], $Q^{AB/F}$ [Ma *et al.*, 2019], PSNR [Wang *et al.*, 2019], SSIM [Wang *et al.*, 2004], and LPIPS [Zhang *et al.*, 2018]. We compared our proposed method with state-of-the-art approaches, including: U2Fusion [Xu *et al.*, 2020], TarDAL [Liu *et al.*, 2022], SwinFusion [Ma *et al.*, 2022], CDDFuse [Zhao *et al.*, 2023a], DDFM [Zhao *et al.*, 2023b], EMMA [Zhao *et al.*, 2024], Text-IF [Yi *et al.*, 2024b] and ControlFusion [Tang *et al.*, 2025].

4.1 Comparisons with other Methods

Existing fusion models are typically developed for clean inputs and thus degrade noticeably when the visible modality suffers from severe corruptions, while most of them also lack a mechanism for instruction-guided control. We therefore benchmark AIR-Fusion under five representative visible degradations (blur, haze, low-light, rain, and noise) and a clean case, covering both in-domain and cross-dataset evaluation. Notably, AIR-Fusion uses one set of parameters to handle all conditions without per-degradation retraining.

ID	Configuration	EN	SD	VIF	$Q^{AB/F}$	PSNR
I	Only noise loss	6.62	40.12	0.90	0.63	28.24
II	No color loss	7.07	44.11	0.94	0.72	30.43
III	No TCR	7.03	43.36	0.87	0.57	29.01
IV	ALL	7.09	44.27	1.04	0.76	29.99

Table 2: Quantitative comparison of ablation experiments.

AIR-Fusion is trained only on LLVIP and tested on LLVIP as well as two unseen datasets (MSRS and RoadScene). For methods without an explicit restoration component, we prepend an off-the-shelf AutoDIR module to restore the degraded visible input before fusion; such pipelines are marked with * in Table 1. For interactive baselines (Text-IF and ControlFusion) and AIR-Fusion, we use degradation-matched prompts, while other methods are run with their default inference settings.

Quantitative Comparisons. Table 1 reports results under degraded conditions. AIR-Fusion achieves the best performance on blur, low-light, and rain across PSNR, SSIM, VIF, and LPIPS, demonstrating strong joint restoration-and-fusion capability with improved perceptual fidelity. For haze, AIR-Fusion attains the highest PSNR and SSIM, the lowest LPIPS, and remains competitive on VIF, indicating better structure recovery and fidelity under scattering. Under noise on the unseen RoadScene dataset, AIR-Fusion achieves the top PSNR and strong VIF, while maintaining competitive LPIPS, suggesting a favorable trade-off between denoising, detail preservation, and perceptual quality. Overall, the consistent performance across datasets supports adapting a frozen restoration-capable diffusion backbone via generative capability elicitation, rather than training a specialized fusion network from scratch.

Qualitative Comparisons. Fig. 3 presents visual comparisons under both degraded inputs and a clean case. Across degradations, AIR-Fusion better suppresses corruption while preserving complementary information from infrared and visible modalities. In blur/rain/haze cases, it restores clearer edges and scene textures with fewer artifacts. Under low-light, it improves visibility with more natural colors and

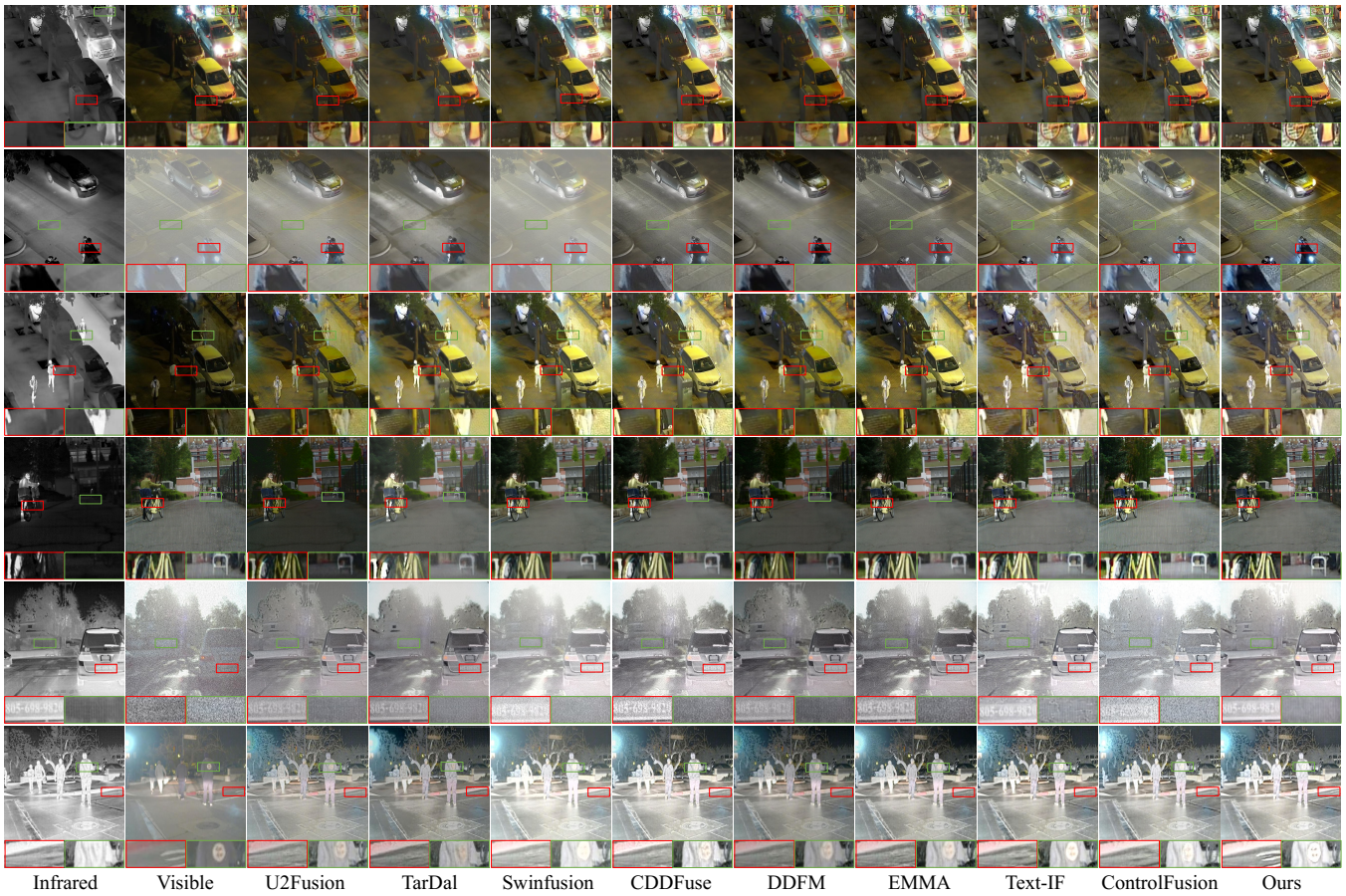

Figure 3: Qualitative comparisons under various visible degradations and a clean case. From top to bottom: blur (LLVIP), haze (LLVIP), low-light (LLVIP), rain (MSRS), noise (RoadScene), and a non-degraded case (RoadScene).

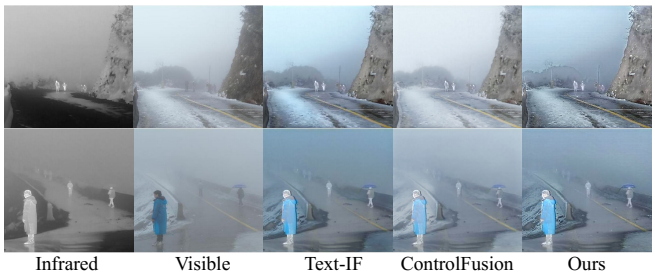

Figure 4: Generalization comparison on real-world haze scenes.

avoids the over-darkening or color shifts observed in competing methods. For noise, AIR-Fusion reduces granular noise while maintaining fine structures. In the clean case (last row), AIR-Fusion highlights thermal targets while retaining visible details such as clothing textures and ground markings, yielding more faithful and visually consistent fusion results.

4.2 Generalization to Real-world Scenarios

Models trained and evaluated on synthetic degradations may still struggle in real-world scenes, where weather patterns and sensor characteristics differ from simulated settings. To assess robustness in the wild, we conduct qualitative evaluation on the public AWMM-100k benchmark [Li *et al.*, 2024], which contains large-scale aligned infrared-visible pairs captured under adverse weather. Specifically, we select real-world haze cases to test transferability beyond the degradations seen in LLVIP training. We compare AIR-Fusion with two instruction-guided baselines, Text-IF [Yi *et al.*, 2024b] and ControlFusion [Tang *et al.*, 2025].

As shown in Fig. 4, other fusion methods trained from scratch exhibit limited transfer to real haze scenes. Text-IF often leaves substantial haze and artifacts, while ControlFusion alleviates haze to some extent but may introduce color shifts and residual veiling. In contrast, AIR-Fusion produces clearer structures and more consistent colors while preserving salient thermal targets from the infrared modality. We attribute this improved transferability to adapting a frozen restoration-capable diffusion backbone with constrained generative adaptation, which provides strong restoration priors beyond the training distribution.

4.3 Performance on High-level Task

To examine whether the fused images preserve task-relevant semantics, we evaluate AIR-Fusion on downstream object detection. Following common practice on LLVIP, we adopt

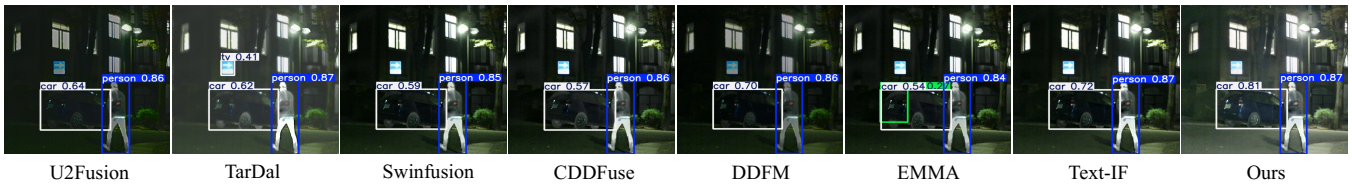

Figure 5: Qualitative comparison of object detection performance.

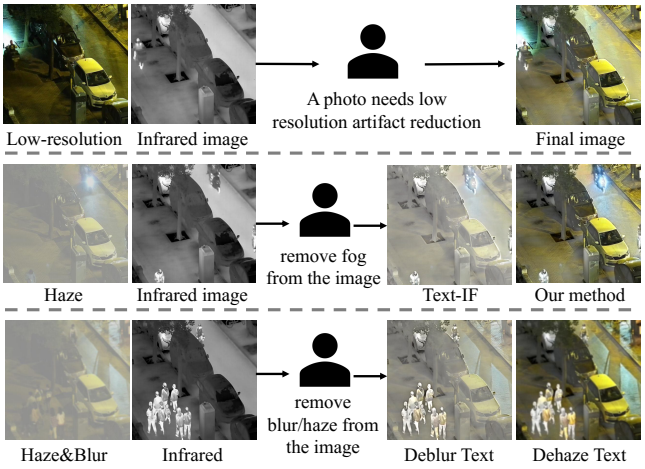

Figure 6: Emergent capabilities of AIR-Fusion. **Top:** low-resolution recovery during fusion. **Middle:** dehazing comparison with Text-IF under text prompts. **Bottom:** prompt-guided selective restoration on compound degradations (Haze&Blur).

YOLOv5 as the detector and fine-tune it on the LLVIP training split. At test time, the detector takes the fused images produced by each fusion method as input. Fig. 5 shows qualitative detection examples. Compared with competing fusion methods, AIR-Fusion yields fused images that lead to more accurate and complete detections, especially for the car and pedestrian under adverse conditions. These results suggest that AIR-Fusion better preserves discriminative structures and salient thermal cues that are beneficial for downstream perception.

4.4 Ablation Study

We ablate key components and report quantitative results in Table 2. Using only diffusion noise loss performs worst because latent denoising alone does not enforce cross-modal complementarity. Removing the color loss can retain competitive reconstruction scores but causes chromatic inconsistency. Disabling TCR hurts structural/perceptual quality, confirming that pixel-space rectification regularizes latent sampling and recovers details. The improved PSNR and structure-related metrics over the No TCR variant further indicate that TCR does not reintroduce degradation artifacts. Overall, the full model performs best by combining fusion-aware supervision with structure-aware rectification.

4.5 Discussion of Emergent Capabilities

Beyond the supervised degraded-IVIF objective, we observe several qualitative behaviors that are not directly targeted by our fusion training data (Fig. 6):

Low-resolution recovery. Although low-resolution degradation is not included in our fusion training set, AIR-Fusion still enhances perceptual sharpness and alleviates blocky pixelation when presented with LR inputs (top). We attribute this behavior to the restoration priors already encoded in the frozen restoration-capable diffusion backbone, which AIR-Fusion can effectively elicit and reuse during fusion through our conditional modulation. This suggests that AIR-Fusion is able to transfer general-purpose restoration capabilities to unseen degradations without task-specific retraining.

Text-guided robustness. This experiment (middle) is designed to compare the language controllability of different methods using a natural human instruction such as "remove fog from the image." Under this prompt, AIR-Fusion responds more faithfully and consistently. In contrast, Text-IF exhibits weaker prompt-following robustness. These observations suggest that AIR-Fusion better aligns the natural-language intent with the restoration behavior during fusion, enabling more reliable instruction-driven editing without drifting away from the fusion objective.

Selective restoration under compound degradations. As shown in the bottom row (e.g., "remove blur/haze from the image"), when prompted with a haze-specific command, AIR-Fusion selectively lifts the global haze veil and restores scene visibility, yet keeps the blur characteristics largely unchanged. Conversely, when using a blur-specific instruction, AIR-Fusion primarily sharpens local structures and recovers fine details, while leaving the atmospheric veil mostly intact instead of over-dehazing. This behavior indicates that AIR-Fusion can disentangle restoration actions via language, enabling fine-grained, instruction-faithful control.

5 Conclusion

We presented AIR-Fusion, a generative capability elicitation framework for degraded infrared–visible image fusion that adapts a frozen restoration-capable latent diffusion backbone via parameter-efficient modules. CMBA aligns infrared cues and textual instructions with the frozen conditioning space for instruction-guided restoration-aware fusion, while TCR constrains denoising with a pixel-latent closed loop to preserve structures and recover fine details. Experiments across datasets and degradations verify consistent gains in fusion fidelity and robustness. Our current evaluation mainly targets visible-side degradations; degraded infrared inputs, dual-modality degradation, and misregistration remain important future directions, together with more efficient inference and broader real-world evaluation.

Acknowledgments

This work was sponsored by the National Natural Science Foundation of China (Nos. 62222608, 62436002, 62476198), the Tianjin Natural Science Funds for Distinguished Young Scholar (No. 23JCJQJC00270), the Zhejiang Provincial Natural Science Foundation of China (No. LD24F020004), and the Natural Science Foundation of Tianjin (No. 25JCYBJC00950).

References

[Bulat *et al.*, 2018] Adrian Bulat, Jing Yang, and Georgios Tzimiropoulos. To learn image super-resolution, use a gan to learn how to do image degradation first. In *Proceedings of the European conference on computer vision (ECCV)*, pages 185–200, 2018.

[Cao *et al.*, 2024] Xuheng Cao, Yusheng Lian, Kaixuan Wang, Chao Ma, and Xianqing Xu. Unsupervised hybrid network of transformer and cnn for blind hyperspectral and multispectral image fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2024.

[Cao *et al.*, 2025] Zihan Cao, Yu Zhong, Ziqi Wang, and Liang-Jian Deng. Mmaif: Multi-task and multi-degradation all-in-one for image fusion with language guidance. *arXiv preprint arXiv:2503.14944*, 2025.

[Chen *et al.*, 2022] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *European conference on computer vision*, pages 17–33. Springer, 2022.

[Chen *et al.*, 2023] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023.

[Fang *et al.*, 2021] Aiqing Fang, Xinbo Zhao, Jiaqi Yang, Beibei Qin, and Yanning Zhang. A light-weight, efficient, and general cross-modal image fusion network. *Neurocomputing*, 463:198–211, 2021.

[Han *et al.*, 2013] 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.

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

[Huang *et al.*, 2023] Ziqi Huang, Kelvin CK Chan, Yuming Jiang, and Ziwei Liu. Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6080–6090, 2023.

[Ibarra-Castanedo *et al.*, 2004] Clemente Ibarra-Castanedo, Daniel Gonzalez, Matthieu Klein, Mariacristina Pilla, Steve Vallerand, and Xavier Maldague. Infrared image processing and data analysis. *Infrared physics & technology*, 46(1-2):75–83, 2004.

[Jia *et al.*, 2021] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3496–3504, 2021.

[Jian *et al.*, 2020] Lihua Jian, Xiaomin Yang, Zheng Liu, Gwanggil Jeon, Mingliang Gao, and David Chisholm. Sedrfuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:1–15, 2020.

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

[Li and Wu, 2024] Hui Li and Xiao-Jun Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103:102147, 2024.

[Li *et al.*, 2024] Xilai Li, Wuyang Liu, Xiaosong Li, Fuqiang Zhou, Huafeng Li, and Feiping Nie. All-weather multi-modality image fusion: Unified framework and 100k benchmark. *arXiv preprint arXiv:2402.02090*, 2024.

[Liu *et al.*, 2017] Yu Liu, Xun Chen, Hu Peng, and Zengfu Wang. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36:191–207, 2017.

[Liu *et al.*, 2022] 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*, pages 5802–5811, 2022.

[Ma *et al.*, 2019] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information fusion*, 45:153–178, 2019.

[Ma *et al.*, 2020] Jiayi Ma, Wei Yu, Chen Chen, Pengwei Liang, Xiaojie Guo, and Junjun Jiang. Pan-gan: An unsupervised pan-sharpening method for remote sensing image fusion. *Information Fusion*, 62:110–120, 2020.

[Ma *et al.*, 2022] 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.

[Mou *et al.*, 2024] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4296–4304, 2024.

[Nichol and Dhariwal, 2021] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021.

[Pradnya and Sachin, 2013] P Mirajkar Pradnya and D Ruikar Sachin. Wavelet based image fusion techniques. In *2013 international conference on intelligent systems and signal processing (ISSP)*, pages 77–81. IEEE, 2013.

[Qi *et al.*, 2024] Chenyang Qi, Zhengzhong Tu, Keren Ye, Mauricio Delbracio, Peyman Milanfar, Qifeng Chen, and Hossein Talebi. Spire: Semantic prompt-driven image restoration. In *European Conference on Computer Vision*, pages 446–464. Springer, 2024.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Shi *et al.*, 2005] Wenzhong Shi, ChangQing Zhu, Yan Tian, and Janet Nichol. Wavelet-based image fusion and quality assessment. *International journal of applied earth observation and geoinformation*, 6(3-4):241–251, 2005.

[Takumi *et al.*, 2017] Karasawa Takumi, Kohei Watanabe, Qishen Ha, Antonio Tejero-De-Pablos, Yoshitaka Ushiku, and Tatsuya Harada. Multispectral object detection for autonomous vehicles. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 35–43, 2017.

[Tang *et al.*, 2022] 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*, 83:79–92, 2022.

[Tang *et al.*, 2023] Linfeng Tang, Xinyu Xiang, Hao Zhang, Meiqi Gong, and Jiayi Ma. Divfusion: Darkness-free infrared and visible image fusion. *Information Fusion*, 91:477–493, 2023.

[Tang *et al.*, 2025] Linfeng Tang, Yeda Wang, Zhanchuan Cai, Junjun Jiang, and Jiayi Ma. Controlfusion: A controllable image fusion network with language-vision degradation prompts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[Wang *et al.*, 2004] 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.

[Wang *et al.*, 2019] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson WH Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12270–12279, 2019.

[Xu *et al.*, 2020] 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, 2020.

[Yi *et al.*, 2024a] Xunpeng Yi, Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110:102450, 2024.

[Yi *et al.*, 2024b] 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*, pages 27026–27035, 2024.

[Zhang and Demiris, 2023] Xingchen Zhang and Yiannis Demiris. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10535–10554, 2023.

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

[Zhang *et al.*, 2020] Yu Zhang, Yu Liu, Peng Sun, Han Yan, Xiaolin Zhao, and Li Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54:99–118, 2020.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[Zhang *et al.*, 2024] Hao Zhang, Lei Cao, and Jiayi Ma. Text-difuse: An interactive multi-modal image fusion framework based on text-modulated diffusion model. *arXiv preprint arXiv:2410.23905*, 2024.

[Zhao *et al.*, 2020] Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Pengfei Li, and Jiangshe Zhang. Didfuse: Deep image decomposition for infrared and visible image fusion. *arXiv preprint arXiv:2003.09210*, 2020.

[Zhao *et al.*, 2023a] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5906–5916, 2023.

[Zhao *et al.*, 2023b] 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*, pages 8082–8093, 2023.

[Zhao *et al.*, 2024] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25912–25921, 2024.

[Zhou *et al.*, 2023] Tao Zhou, Qi Li, Huiling Lu, Qianru Cheng, and Xiangxiang Zhang. Gan review: Models and medical image fusion applications. *Information Fusion*, 91:134–148, 2023.