GRANFLOW: GRANULARITY-MATCHED CONDITIONING FOR DEGRADATION-ROBUST INFRARED-VISIBLE FUSION VIA RECTIFIED FLOW

Anonymous authors
Paper under double-blind review

ABSTRACT

Compound degradations in real-world infrared–visible imaging undermine the reliable integration of complementary information, making degradation-robust fusion essential for preserving scene structure and restoring visual appearance. These degradations affect image content in fundamentally different ways: noise and stripe artifacts introduce spurious artifacts, blur attenuates high-frequency detail, and haze induces spatially varying contrast loss together with broader shifts in image statistics. Distinguishing frequency-dependent corruption from broader distributional shifts can therefore help balance artifact suppression, complementary detail preservation, and global appearance restoration. We propose GranFlow, a Rectified Flow fusion framework combining a wavelet-domain gated module that adaptively reweights modalities for frequency-selective corruptions, and a scale-aware global conditioning mechanism for distributional shifts. We find that this global signal captures sample-dependent variation in latent displacement scale, producing scale-adaptive modulation without degradation labels. Experiments across single and compound degradations demonstrate that degradation-adaptive modality reweighting and scale-adaptive latent refinement improve structural fidelity and visual appearance while preserving clean-input fusion quality and benefiting downstream object detection.

1 INTRODUCTION

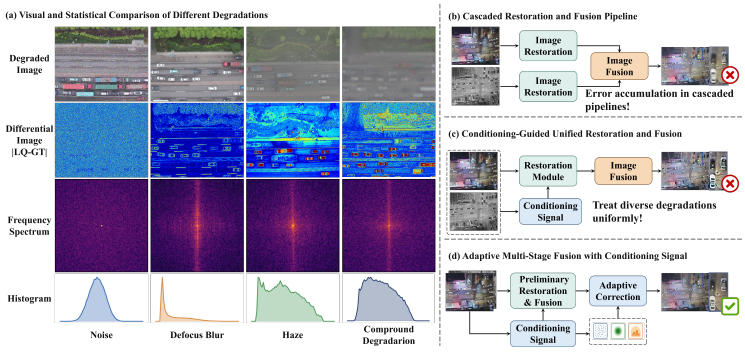

Figure 1: Frequency-domain degradation signatures (left) and pipeline comparisons (right).

Infrared-visible image fusion aims to integrate complementary information from thermal and visible spectra, producing a single image that preserves salient targets under low-light or adverse conditions while retaining rich visible-domain texture. This capability underpins applications including nighttime surveillance, autonomous driving, and remote sensing (Cao et al., 2023; Liu et al., 2024). To understand how degradations affect fusion inputs, Figure 1(a) compares noise, defocus blur, haze and compound degradation in image, residual, frequency and histogram domains. These visual and statistical observations reveal that the degradations differ not only in severity but also in their underlying corruption mechanisms. Noise mainly introduces random spatial variations and

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

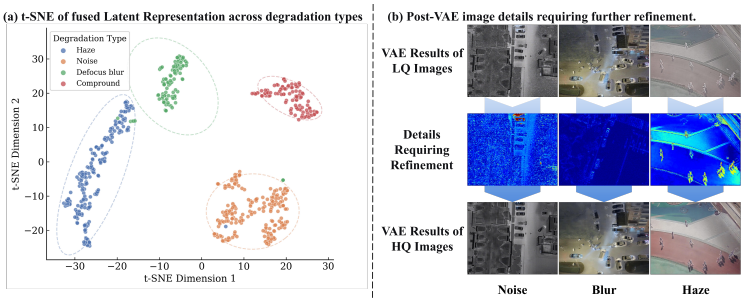

Figure 2: (a) t-SNE of fused latent representations across degradation types, showing degradation-related grouping. (b) Post-VAE results of low-quality and high-quality inputs, with difference maps highlighting the details that remain to be refined.

unwanted high-frequency components; denoising should therefore suppress these components. Defocus blur attenuates high frequencies and removes fine structures, requiring detail reconstruction or supplementation. Haze creates a spatially smooth veiling effect with global contrast and intensity-distribution shifts, which cannot be adequately modeled as a localized frequency anomaly. More importantly, compound degradation combines multiple signatures across the residual, frequency, and intensity domains. Since its constituent degradations may interact, the resulting corruption is more complex than any individual degradation and difficult to address with a single correction mechanism.

Figure 1(b)-(d) further illustrate how these observations motivate our restoration and fusion design. The cascaded pipeline in Figure 1(b) restores the source images before fusion, causing restoration errors to propagate into and potentially be amplified by the fusion stage. The unified framework in Figure 1(c) reduces this separation through conditioning-guided joint restoration and fusion. However, its uniform conditioning cannot accommodate the distinct residual, frequency, and distributional characteristics of different degradations, limiting both degradation-specific correction for individual corruptions and the coordination of complementary operations for compound degradation. To overcome these limitations, Figure 1(d) introduces an adaptive multi-stage strategy in which degradation-aware conditioning signals guide both preliminary restoration and fusion and subsequent refinement. The initial stage produces a coarsely restored and fused result, while the following stages apply complementary, degradation-specific corrections according to the identified corruption characteristics. This design enables differentiated and coordinated treatment of heterogeneous and compound degradations while mitigating the error accumulation inherent in conventional cascaded pipelines.

Following this principle, we propose GranFlow, a two-stage Rectified Flow framework that implements the strategy in Figure 1(d) through two conditioning mechanisms at matched granularities. Stage I addresses spatially localized and frequency-selective corruptions, such as noise and blur-induced high-frequency loss. A wavelet-domain degradation-aware module extracts subband statistics from each modality and drives gated fusion, reweighting modality contributions at the pixel and channel levels according to local degradation severity. Beyond preliminary restoration, Stage I also organizes the latent space by degradation characteristics. As shown in Figure 2(a), Noise, Defocus blur, and Haze form distinct distributions, while Compound is clearly separated from the other types due to their joint effects. This indicates that the encoder preserves degradation-aware cues that can serve as conditioning information without explicit degradation labels or type-specific branches. However, latent separability alone does not guarantee faithful reconstruction. Figure 2(b) shows that low- and high-quality inputs produce degradation-specific residuals after the VAE, including residual noise, unrecovered fine structures under blur, and highlighted boundaries with global color and contrast shifts under haze. These residuals reflect sample-dependent distributional discrepancies rather than localized frequency anomalies. We therefore introduce a scale-aware mechanism in Stage II to measure latent-space discrepancies and inject them through scale-gated normalization, thereby modulating the Rectified Flow trajectory. By encoding the sample-dependent latent displacement, the scale embedding controls the magnitude, rather than the type, of correction and complements the static conditional summary from Stage I. Our technical contributions are threefold:

- We propose GranFlow, a two-stage Rectified Flow framework for degradation-robust infrared-visible fusion, which decouples degradation-adaptive pixel-level restoration from sample-adaptive latent refinement so that each correction matches the granularity of its target degradation, instead of a single uniform conditioning signal.
- We introduce two complementary conditioning mechanisms: a wavelet module encodes modalities into subband descriptors and drives a dense gate, reweighting pixel contributions by subband degradation evidence to suppress artifacts, restore blurred detail, and separate latent cues. A scale-aware mechanism that injects sample-dependent latent displacement via scale-gated normalization, correcting the residual contrast and color deviation left by localized filtering.
- We show that pairing a static, spatially resolved signal with a trajectory-dependent, sample-level one lets one model separate frequency-dependent corruption from broader distributional shifts, improving restoration-fusion performance without degradation labels or type-specific branches.

2 Related Work

2.1 Infrared-Visible Image Fusion with Clean Inputs

Infrared-visible image fusion integrates complementary thermal and visible information into a single image preserving salient targets and texture details. Early autoencoder methods learn a fusion-friendly latent space via reconstruction, but leave the fusion rule as a separate post hoc step (Li & Wu, 2018; Zhao et al., 2020). End-to-end learnable fusion closes this gap via adversarial distribution matching (Ma et al., 2019) or direct supervision with structural and perceptual losses (Sun et al., 2022b). Since convolutions have limited receptive fields, Transformer-based methods (Wang et al., 2022; Tang et al., 2022) address long-range cross-modal dependency, while correlation-driven decomposition methods (Zhao et al., 2023a; 2024) explicitly separate modality-shared and modality-specific components. More recently, generative modeling has been adopted as a stable alternative to adversarial training (Ho et al., 2020; Song et al., 2020): DDFM (Zhao et al., 2023b) and Diff-IF (Yi et al., 2024a) establish the feasibility of diffusion-based fusion, and Rectified Flow and flow matching (Liu et al., 2022; Lipman et al., 2022) further reduce sampling cost, as in RF-Fusion (Wang et al., 2026). Despite their differing architectures and sampling schemes, these methods share two assumptions: that both inputs are clean, and that a global separation between shared and specific components is well defined. Both break down under complex conditions, where different modalities may lose reliable content in different regions, leaving no consistent basis for such a separation.

2.2 Infrared-Visible Image Fusion under Degraded Inputs

To relax the clean-input assumption, subsequent work introduces restoration into the fusion pipeline, yet fragments along degradation types. One group handles weather-induced corruption (rain, snow, haze) via physical priors or weather-specific prompts (Chen et al., 2025; Li et al., 2026a;b; 2025b; Zhang et al., 2026b), limiting generalization when corruption mismatches. A second branch targets sensor- or motion-induced degradation (blur, low resolution, noise) via dedicated sub-networks (Tang et al., 2024; Xue et al., 2024; Zhang et al., 2024; Shi et al., 2026), assuming the degradation type is known a priori. A third direction handles both within a unified framework via a shared estimation module (attention map, prompt, or conditioning embedding) applied uniformly regardless of degradation origin (Zhang et al., 2026a; Yi et al., 2024b; Cao et al., 2025; Tang et al., 2026a; Wang et al., 2025). Generative approaches follow the same pattern: degradation-aware diffusion methods rely either on a single global signal (gating strength, observation constraint, or text embedding) applied uniformly (Yang et al., 2025; Zhang et al., 2025), or on a degradation type fixed at training time rather than discriminated at inference (Li et al., 2025a), while generic conditioning tools such as classifier-free guidance (Ho & Salimans, 2022) and AdaIN (Huang & Belongie, 2017) likewise inject a single global signal. Dynamic mixture-of-experts paradigms (Sun et al., 2025) or adaptive discrimination (Chen et al., 2026) likewise retain a unified structure rather than explicitly matching degradation mechanisms to spatial-frequency granularity. Across all these branches, the corrective mechanism rarely adapts to the spatial and frequency characteristics of the specific degradation, so correction cannot vary spatially or spectrally. This motivates treating degradation

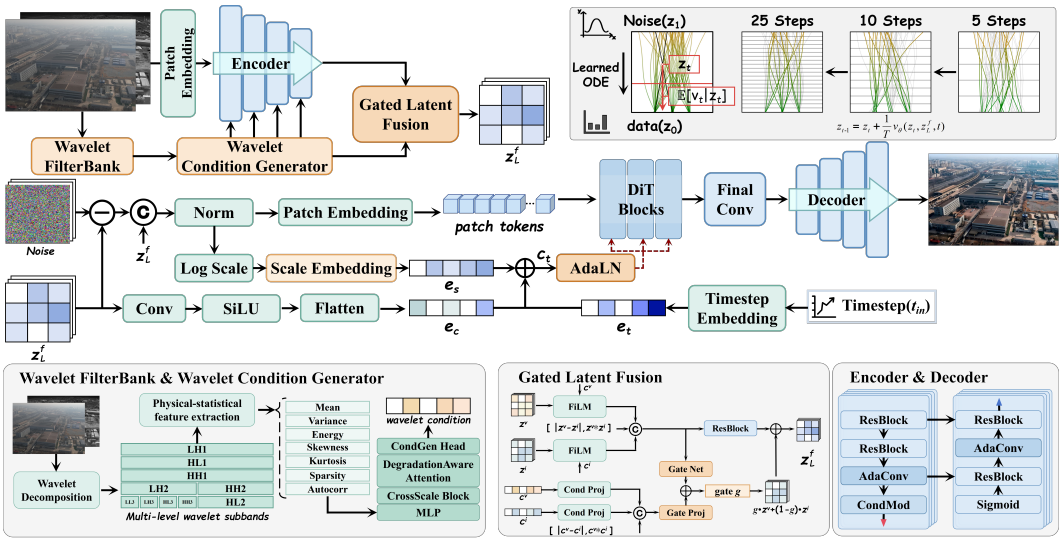

Figure 3: **Overview of GranFlow.** The framework comprises two stages: (I) a wavelet-conditioned dual-branch VAE for per-modality restoration and latent fusion, and (II) a conditional DiT that refines the fused latent through Rectified Flow, decoded by the frozen Stage-I decoder.

granularity and physical mechanisms, not merely its presence, as a first-class design consideration. Our work builds on Rectified Flow's efficiency but departs from this global-or-homogeneous conditioning paradigm by explicitly disentangling degradation conditioning at matched granularities within the flow trajectory.

3 Method

3.1 Overall Framework

Let (x_L^v, x_L^i) and (x_G^v, x_G^i) denote degraded and clean visible-infrared pairs. GranFlow consists of two stages. Stage I employs a dual-branch VAE with wavelet-conditioned encoders. The resulting modality latents and skip features are adaptively fused into z_L^f and h_L^f, producing a preliminary restoration-fusion result. In Stage II, the autoencoder is frozen and a conditional DiT predicts the Rectified Flow velocity from z_L^f toward z_G^f. It receives timestep information, a trajectory-independent summary of z_L^f, and a scale embedding derived from the current latent displacement. At inference, only degraded inputs are required; 25 reverse Euler steps generate the refined latent, which is decoded using the frozen decoder and Stage-I skip features.

3.2 Stage-I: Dual-Branch Variational AutoEncoder with Wavelet Conditioning

Visible and infrared inputs are encoded by separate branches sharing one architecture. For $m \in \{v, i\}$:

$$(z^m, h^m, c^m, \mu^m, \ell^m) = E_m(x^m), \tag{1}$$

where $z^m \in \mathbb{R}^{B\times C\times H\times W}$ is the latent, h^m are shallow skip features, $c^m \in \mathbb{R}^{512}$ is the wavelet condition, and (μ^m, ℓ^m) are VAE posterior parameters with $z^m = \mu^m + \exp(\ell^m/2) \odot \epsilon$ during training. The decoder maps the fused latent back through a symmetric convolutional path, incorporates fused skip features, and outputs a 3-channel image.

3.2.1 Wavelet FilterBank & Wavelet Condition Generator

We compute, prior to any restoration, an explicit degradation-dependent frequency summary $c^m = C_m(x_L^m)$ for each modality (Fig. 3, bottom-left), which is then broadcast as a conditioning signal to the downstream modules.

A three-level filter bank decomposes x_L^m into ten subbands (LL3 plus LH/HL/HH per level). Each subband is encoded by a 64-D descriptor: a degradation-sensitive 7-D closed-form statistics vector (mean, variance, energy, skewness, kurtosis, sparsity, autocorrelation) plus a 57-D CNN feature preserving local spatial evidence. This yields a 10×64 tensor per modality, whose subband distribution reveals degradation type. These descriptors pass through a degradation-aware attention module and a lightweight cross-scale block that models cross-frequency dependencies and resolves inter-subband ambiguity. The flattened 640-D vector is then linearly projected to $c^m \in \mathbb{R}^{512}$. This global degradation-aware signal modulates per-layer restoratio

3.2.2 ENCODER & DECODER

Each encoder block refines features through two complementary adaptation mechanisms: AdaConv and CondMod. Each decoder block uses only AdaConv, while residual fusion combines dual-modality latents.

AdaConv & CondMod. AdaConv is applied in both encoder and decoder blocks; it adjusts channel-mixing behavior according to the feature's own implicit degradation statistics. CondMod is used only in the encoder: it injects the explicit wavelet condition c^m via residual cross-attention to selectively emphasize condition-relevant channels. Full formulations are provided in Appendix A.2.

Condition-Aware Gated Latent Fusion. Let κ, ρ be shared learnable scalars. Given latents z^v, z^i, fusion applies condition-specific residual FiLM,

$$\hat{z}^m = z^m \odot (1 + \kappa \tanh \gamma^m) + \kappa \tanh \beta^m, \qquad [\gamma^m, \beta^m] = P_m(c^m), \tag{2}$$

then predicts a spatial-channel gate from the latent interaction

$$q = [\hat{z}^v, \hat{z}^i, |\hat{z}^v - \hat{z}^i|, \hat{z}^v \odot \hat{z}^i] \tag{3}$$

and a paired-condition bias $b^p = P_g([c^v, c^i, |c^v - c^i|, c^v \odot c^i])$:

$$g = \sigma(G(q) + \kappa \tanh b^p), \tag{4}$$

$$z^f_{\text{mix}} = g \odot \hat{z}^v + (1 - g) \odot \hat{z}^i. \tag{5}$$

A residual refinement and GroupNorm generate

$$z^f = \text{GN}(z^f_{\text{mix}} + \rho R(q)), \tag{6}$$

this lets the network select modality contributions per-pixel and per-channel from paired degradation evidence, rather than relying on fixed averaging.

Stage-I Training Objective. Stage I is trained with Charbonnier self-supervised modality reconstruction and fusion-consistency losses over direct and cross latent/skip combinations, with wavelet-fusion, gradient and KL:

$$\mathcal{L}_{\text{AE}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{wave}} + \mathcal{L}_{\text{edge}} + \mathcal{L}_{\text{KL}}. \tag{7}$$

3.3 STAGE-II: LATENT RECTIFIED FLOW WITH FROZEN AUTOENCODER

With Stage-I parameters ϕ frozen, deterministic latents z_L^f, z_G^f are formed as in Stage I using μ^m in place of sampled z^m. We adopt a Rectified Flow formulation that linearly interpolates between the degraded endpoint ($t = T$) and the clean endpoint ($t \to 0$),

$$z_t = \lambda_t z_L^f + (1 - \lambda_t) z_G^f + \eta_t, \qquad \lambda_t = t/T, \quad t \in \{1, \ldots, T = 25\}, \tag{8}$$

with a small training-time perturbation η_t that vanishes at both endpoints and peaks at mid-trajectory, encouraging robustness away from the exact linear path. A DiT-based backbone predicts the RF velocity $v_\theta(z_t, z_L^f, t)$, conditioned on the degraded latent z_L^f and a shared conditioning vector

$$c_t = e_t + s_c + r_t, \tag{9}$$

combining the timestep embedding e_t with the two signals below.

Let $\delta_t = z_t - z_L^f$ denote the residual and $s_t = \text{std}(\delta_t)$ its per-sample magnitude. Rather than feeding z_t directly, the backbone takes

$$\tilde{\delta}_t = (1 - \alpha_t)\, \delta_t + \alpha_t \, \frac{\delta_t}{\sqrt{s_t^2 + \epsilon_0^2}}, \qquad \alpha_t = \text{smoothstep}(s_t; \tau_1, \tau_2), \tag{10}$$

which retains the native scale of easy samples while normalizing hard ones, thereby decoupling the *direction* of the correction, carried by the input channels, from its *magnitude*, carried by r_t below. The network input is $[\tilde{\delta}_t,\ z_L^f]$.

3.3.1 SCALE EMBEDDING

To expose sample-dependent residual scale beyond the timestep, we embed the residual magnitude s_t defined above as

$$r_t = \mathcal{MLP}(\log(1 + s_t)), \tag{11}$$

via a 2-layer MLP. On the ideal path $s_t \approx (1 - \lambda_t)d$ with $d = \mathrm{std}(z_G^f - z_L^f)$, so s_t couples the schedule with a sample-dependent latent displacement scale d that the timestep alone cannot express; at inference the trajectory departs from this path and s_t instead reports the current displacement scale along the predicted trajectory. It is also non-redundant by construction: the residual magnitude is factored out of the input channels by the scale-gated normalization above, leaving the conditioning path as the only route to recover it. Since $s_T = 0$ at the start, r_T reduces to a learned constant and the first step is driven by s_c; from $t < T$ onward r_t acts as a global correction bias matched to the sample's own displacement scale.

3.3.2 CONDITIONAL SUMMARY

In parallel, a lightweight convolutional path summarizes z_L^f into a global vector $s_c = \mathcal{S}(z_L^f)$. Unlike r_t, s_c is trajectory-independent: it is available at every step, including $t = T$ where the residual is identically zero, and therefore carries the degradation prior that initiates the first Euler step. It also supplies each DiT block with global scene context beyond the patch-local tokens.

3.3.3 RECTIFIED FLOW OBJECTIVE AND INFERENCE

The target velocity is $v^\star = z_G^f - z_L^f$, with predicted endpoint $\hat{z}_G^f = z_L^f + v_\theta(z_t, z_L^f, t)$. The loss combines a norm-weighted velocity MSE, an endpoint MSE (upweighted for $t > T/4$) and a one-step consistency term against the ideal interpolated state $z_{t-1}^\star$:

$$\mathcal{L}_{\mathrm{RF}} = \mathcal{L}_{\mathrm{vel}} + \mathcal{L}_{x_0} + \mathcal{L}_{\mathrm{step}} \tag{12}$$

At inference, starting from $z^{(T)} = z_L^f$, 25 reverse Euler steps

$$z^{(t-1)} = z^{(t)} + \frac{1}{25}\, v_\theta(z^{(t)}, z_L^f, t) \tag{13}$$

produce $z^{(0)}$, decoded with the frozen decoder and low-quality(LQ) skip features h_L^f into the final fusion result $\hat{y}$.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTING

Implementation Details. All experiments are conducted on four NVIDIA GeForce RTX 4090 GPUs, with the model implemented in PyTorch 2.1.0. The model training is conducted in two stages. In the first stage, the VAE component is trained, for which we employ the Lion optimizer with an initial learning rate of 3.0×10^{-5}, scheduled via a cosine annealing decay. Input images are randomly cropped to 256×256 pixels. In the second stage, the Latent Rectified Flow component is trained, where we adopt the AdamW optimizer with an initial learning rate of 1.0×10^{-4}, also adjusted by a cosine annealing schedule. Owing to space constraints, the parameter study on the number of Latent Rectified Flow steps and the comparison of computational efficiency are deferred to the supplementary material A.10 and A.11.

Dataset Protocol. We train on DroneVehicle (Sun et al., 2022a), DVTOD (Song et al., 2024), M3SVD (Tang et al., 2026b), and LLVIP (Jia et al., 2021), with all RGB-IR pairs standardized to 1280×1024. A stochastic, modality-aware degradation pipeline synthesizes diverse combinations of noise, blur, stripe artifacts, and haze; implementation details are provided in Appendix A.3. Evaluation is conducted on DeMMI-RF (Sun et al., 2025) and EMS (Yi et al., 2024b), supplemented by the

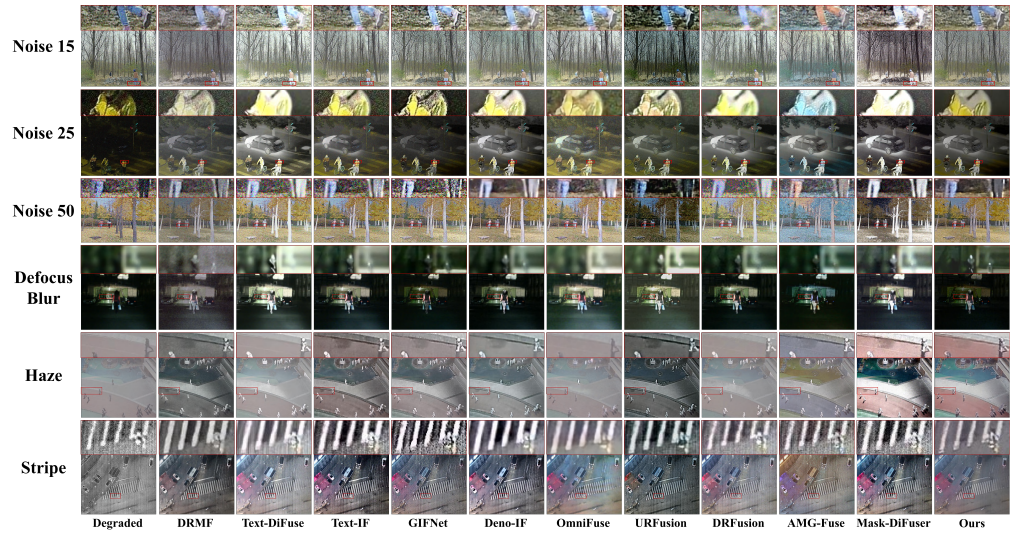

Figure 4: Visual comparison of fusion results under different degradation types on DeMMI-RF.

real-world AWMM (Li et al., 2026b) and VisDrone (Zhu et al., 2021) datasets, as well as extreme-scene subsets curated from DVTOD and M3SVD, to assess robustness under severe, compounded, and naturally occurring degradations. We further evaluate downstream robustness via object detection on DroneVehicle and LLVIP under defocus blur, and semantic segmentation on FMB (Liu et al., 2023) under haze.

Competing Methods. We compare with ten recent approaches, namely DRMF (Tang et al., 2024), Text-DiFuse (Zhang et al., 2024), Text-IF (Yi et al., 2024b), GIFNet (Cheng et al., 2025), Deno-IF (Xu et al., 2026), OmniFuse (Zhang et al., 2025), URFusion (Xu et al., 2025), DRFusion (Shi et al., 2026), AMG-Fuse (Li et al., 2026a), and Mask-DiFuser (Tang et al., 2025). In addition to these ten methods, we include the Rectified Flow-based fusion baseline RF-Fusion (Wang et al., 2026) in a dedicated comparison in Appendix A.7. We adopt six evaluation metrics (Wang et al., 2008), namely QNCIE, CC, SCD, MSE, PSNR, and MS-SSIM. For each metric, the best, second-best, and third-best results are denoted in Red, ForestGreen, and Blue, respectively.

4.2 All-in-One Degraded Image Fusion Results

Fig. 4 shows fusion results under various degradations (noise, defocus blur, haze, stripe) on DeMMI-RF; EMS results are provided in Appendix A.4. Our method produces clearer structures with fewer artifacts, effectively suppressing noise and stripe interference while preserving edges and textures. For blurred and hazy scenes, it recovers recognizable object contours. In contrast, several competing methods introduce over-smoothing, color distortion, or residual degradation.

Quantitative comparisons are reported in Table 1. GranFlow ranks first on all six metrics on DeMMI-RF and on five of six on EMS (CC excepted), exceeding the second-best method by 4.34% and 4.59% on average, respectively. These results are consistent with the intended balance between suppressing spurious frequency components and preserving complementary structural detail through frequency-aware modality reweighting.

4.3 Multiple Degradation, Degradation-Free and Real-World Degradation Scenario Results

Fig. 5, Table 2 and Table 3 cover compound, degradation-free, and real-world scenarios. GranFlow leads on MSE/PSNR/SCD under Haze+Blur, on all six metrics under the four-fold Haze+Blur+Noise+Stripe, and on five of six in the degradation-free case. These results support the effectiveness of pairing degradation-aware modality selection with scale-adaptive latent refinement, while the clean-input results indicate that robustness does not come at the expense of fusion quality.

Table 1: Comparison of fusion methods on the DeMMI-RF and EMS datasets.

Method	DeMMI-RF						EMS					
	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM
DRMF	0.8778	0.6858	1.7246	0.0364	14.9199	0.6659	0.8721	0.5553	1.8560	0.0599	12.4603	0.6060
Text-DiFuse	0.8732	0.6459	1.2501	0.0775	11.6745	0.6408	0.8596	0.4586	1.2519	0.1157	9.6005	0.5073
Text-IF	0.8751	0.6381	1.4467	0.0466	14.0862	0.6338	0.8712	0.5089	1.6072	0.0714	11.6704	0.5950
GIFNet	0.8700	0.6463	1.4181	0.0417	14.4375	0.6211	0.8706	0.5382	1.6382	0.0527	12.9350	0.6005
Deno-IF	0.8857	0.5803	1.5047	0.0723	13.4977	0.6291	0.8696	0.5208	1.7434	0.0591	12.4418	0.5728
OmniFuse	0.8716	0.6362	1.5562	0.0562	13.3876	0.6570	0.8716	0.4837	1.6236	0.0773	11.2677	0.5967
URFusion	0.8781	0.6692	1.4980	0.0476	13.5931	0.6347	0.8748	0.5036	1.3344	0.0838	11.1454	0.5669
DRFusion	0.8769	0.6324	1.4233	0.0531	13.6236	0.6619	0.8671	0.5072	1.5875	0.0722	11.6001	0.6023
AMG-Fuse	0.8786	0.6440	1.4632	0.0507	14.0200	0.6672	0.8687	0.5105	1.6174	0.0589	12.4797	0.6031
Mask-DiFuser	0.8843	0.6953	1.0150	0.0583	12.6846	0.6226	0.8741	0.5633	1.2630	0.0722	11.6674	0.5755
GranFlow(Ours)	0.8887	0.7030	1.8639	0.0311	16.0177	0.7344	0.8761	0.5450	1.8975	0.0468	13.4665	0.6379

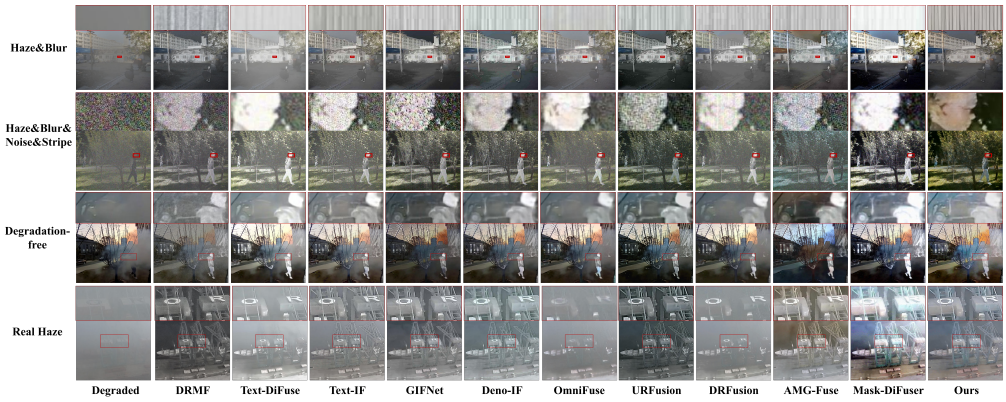

Figure 5: Qualitative results under compound degradation, degradation-free, and real-world scenarios.

4.4 Ablation Study

Table 4 evaluates the effects of the proposed conditioning granularities and adaptive modules. Removing wavelet conditioning degrades the fusion performance, demonstrating the importance of frequency-aware guidance for preliminary restoration and fusion. The removal of either AdaConv or CondMod also causes performance deterioration, while replacing gated fusion with simple weighted averaging weakens adaptive modality selection. In the latent refinement stage, removing the scale embedding further reduces performance, confirming the importance of trajectory-dependent displacement information. Notably, s_c and r_t achieve their best effect when used jointly: relying on either signal alone increases the dependency burden on the remaining condition and may lead to imbalanced conditioning. Their complementary interaction therefore provides more effective guidance throughout the flow trajectory, consistently improving the evaluated metrics over timestep-only conditioning.

Table 2: Quantitative comparison under compound degradations on DVTOD and M3SVD datasets.

Method	Haze+Blur						Haze+Blur+Noise+Stripe					
	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM
DRMF	0.8746	0.5864	1.8148	0.0423	14.3439	0.6285	0.8377	0.4051	1.4795	0.0487	13.3705	0.4395
Text-DiFuse	0.8582	0.5383	1.6632	0.0738	11.7381	0.6418	0.8497	0.4880	1.6026	0.0984	10.3144	0.4313
Text-IF	0.8598	0.5531	1.7428	0.0451	14.1927	0.6583	0.8255	0.3213	1.3666	0.0675	12.0845	0.3610
GIFNet	0.8828	0.5683	1.8352	0.0379	14.8837	0.6798	0.8291	0.3302	1.1263	0.0622	12.3556	0.3030
Deno-IF	0.8773	0.5748	1.7990	0.0405	14.6309	0.6735	0.8611	0.5310	1.7237	0.0392	14.5565	0.5307
OmniFuse	0.8598	0.5365	1.7236	0.0531	13.5483	0.6314	0.8471	0.4673	1.6473	0.0521	13.3265	0.4642
URFusion	0.8706	0.5941	1.8008	0.0418	14.3729	0.6798	0.8440	0.4742	1.5902	0.0456	13.7414	0.3870
DRFusion	0.8595	0.5338	1.7208	0.0521	13.6887	0.6207	0.8433	0.4411	1.6025	0.0550	13.1450	0.4204
AMG-Fuse	0.8710	0.5478	1.6700	0.0418	14.4759	0.6676	0.8273	0.3377	1.3295	0.0525	13.1752	0.3344
Mask-DiFuser	0.8722	0.6129	1.3870	0.0526	13.3800	0.6683	0.8653	0.5545	1.3950	0.0480	13.5896	0.3956
GranFlow(Ours)	0.8739	0.6088	1.8353	0.0355	15.2436	0.6793	0.8682	0.5827	1.8550	0.0351	15.1564	0.6301

Table 3: Degradation-free fusion results on M3SVD dataset.

Method	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM
DRMF	0.8745	0.5644	1.9232	0.0441	13.9723	0.6294
Text-DiFuse	0.8740	0.5116	1.1843	0.1152	9.6129	0.6159
Text-IF	0.8831	0.4952	1.5944	0.0804	11.4414	0.6419
GIFNet	0.8726	0.5208	1.7646	0.0409	14.7478	0.6713
Deno-IF	0.8818	0.5442	1.7799	0.0532	13.6248	0.6699
OmniFuse	0.8802	0.4884	1.5426	0.0764	11.7033	0.6365
URFusion	0.8788	0.4922	1.2910	0.0596	12.6626	0.6201
DRFusion	0.8771	0.5019	1.5629	0.0678	12.6023	0.6454
AMG-Fuse	0.8693	0.4967	1.5508	0.0525	13.6867	0.6302
Mask-DiFuser	0.8781	0.5690	1.3241	0.0602	12.8088	0.6442
GranFlow(Ours)	0.8833	0.5457	1.9561	0.0370	15.2704	0.6731

Table 4: Ablation study on DeMMI-RF dataset.

Method	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM
VAE_baseline	0.8791	0.6751	1.7494	0.0336	15.5496	0.6873
w/o wavelet cond	0.8878	0.6973	1.7837	0.0336	15.5543	0.7146
w/o AdaConv	0.8845	0.6939	1.7930	0.0316	15.8488	0.7123
w/o CondMod	0.8830	0.6812	1.7878	0.0316	15.8638	0.7005
w/o GatedLatentFusion	0.8862	0.6981	1.8233	0.0312	15.9731	0.7258
TimeStep Embedding only	0.8857	0.7010	1.7955	0.0315	15.8790	0.7153
w/o Condition Summary	0.8833	0.6873	1.7668	0.0336	15.5304	0.6981
w/o Scale Embedding	0.8847	0.6965	1.7972	0.0329	15.6494	0.7154
Full	0.8887	0.7030	1.8639	0.0311	16.0177	0.7344

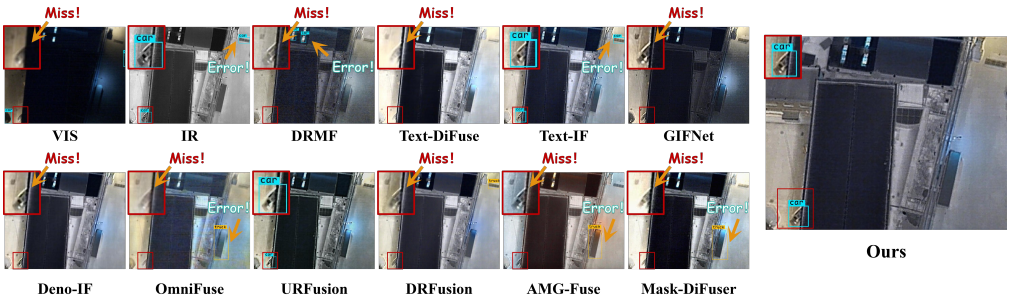

Figure 6: Qualitative object detection results on the DroneVehicle dataset.

4.5 Downstream Task Experiments

To verify the effectiveness of our method on high-level semantic tasks, we evaluate object detection performance on the blurry DroneVehicle and LLVIP datasets using a fixed YOLOv11 (Khanam & Hussain, 2024) detector. Table 5 shows that the images generated by our method improve the detector's recognition ability for blurry or partially occluded targets; the case in Figure 6 further shows that the method better preserves target contours and local structures usable for localization. GranFlow achieves the highest mAP50-95 on both datasets, showing that adaptive restoration and fusion preserves information useful for object localization rather than merely improving visual appearance. Additional qualitative results are provided in Appendix A.8 and A.9.

Table 5: Comparison of object detection results across different methods on DroneVehicle and LLVIP datasets.

Dataset	Metric	VI	IR	DRMF	Text-DiFuse	Text-IF	GIFNet	Deno-IF	OmniFuse	URFusion	DRFusion	AMG-Fuse	Mask-DiFuser	GranFlow
DroneVehicle	mAP50	0.7710	0.7861	0.7913	0.8039	0.8175	0.8138	0.4575	0.8085	0.8316	0.7980	0.8159	0.7955	0.8318
	mAP50-95	0.4838	0.5564	0.5467	0.5436	0.5758	0.5702	0.3206	0.5331	0.5808	0.5526	0.5787	0.5565	0.5872
	precision	0.7848	0.7837	0.7575	0.8057	0.8243	0.8025	0.6964	0.7918	0.7985	0.7940	0.8039	0.7517	0.8390
	recall	0.7339	0.7336	0.7596	0.7497	0.7645	0.7776	0.4518	0.7724	0.7842	0.7544	0.7607	0.7577	0.7809
LLVIP	mAP50	0.9050	0.9640	0.9600	0.9650	0.9640	0.9670	0.9150	0.9630	0.9580	0.9650	0.9660	0.9580	0.9690
	mAP50-95	0.5000	0.6310	0.6050	0.6130	0.6210	0.6190	0.5800	0.5980	0.6110	0.6180	0.6140	0.6100	0.6510
	precision	0.9060	0.9230	0.9270	0.9300	0.9340	0.9200	0.9160	0.9520	0.9330	0.9310	0.9220	0.9290	0.9500
	recall	0.8370	0.9220	0.9040	0.9240	0.9020	0.9170	0.8770	0.8900	0.9080	0.9180	0.9160	0.9050	0.9370

5 Conclusion

In this work, we presented GranFlow, a latent Rectified Flow framework that treats degradation robustness as an adaptive information-allocation problem rather than a uniform restoration task. By matching conditioning granularity to degradation characteristics, GranFlow adaptively determines which modality to trust locally and how strongly to refine each sample globally, using wavelet-conditioned spatial and channel gating and trajectory-aware scale conditioning. Results across diverse single, compound, clean, and real-world scenarios show that this design improves structural fidelity and visual quality without degradation labels or degradation-specific branches. Its consistent detection gains on DroneVehicle and LLVIP further support the use of GranFlow as a robust fusion front end for low-altitude UAV inspection and nighttime monitoring.

AI USE STATEMENT

In this work, we used generative AI tools for code debugging and text polishing. Specifically, we used AI assistants to help identify and correct errors in parts of the PyTorch implementation, and to improve the phrasing and readability of some paragraphs in the manuscript. We have not used generative AI tools for generating research ideas, designing experiments, analyzing data, or writing the full manuscript. All AI-assisted code was manually reviewed, tested on representative inputs, and verified by at least two authors to ensure correctness and alignment with the method described in the paper. All AI-polished text was carefully read and edited by the authors; we take full responsibility for the final content of this work, including any claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work focuses on infrared-visible image fusion under degraded conditions, using publicly available benchmark datasets (DroneVehicle, DVTOD, M3SVD, LLVIP, DeMMI-RF, EMS, and FMB). The research does not involve human subjects, animal experiments, or the collection of personally identifiable information. All datasets used are intended for academic research and are cited appropriately. While the proposed method could be applied in surveillance or autonomous driving systems, this paper only addresses the technical problem of improving fusion quality under degradation; we do not release any deployment-ready surveillance system or advocate for specific monitoring applications. The authors declare no conflicts of interest or sponsorship that could have influenced the design, execution, or reporting of this study.

REPRODUCIBILITY STATEMENT

We make substantial efforts to ensure the reproducibility of our results. The complete architecture of GranFlow, including the dual-branch VAE with wavelet conditioning, AdaConv, CondMod, and the latent Rectified Flow stage, is described in Section 3, with additional module details and formulas provided in Appendices A.2 and A.3. Training datasets (DroneVehicle, DVTOD, M3SVD, LLVIP) and evaluation datasets (DeMMI-RF, EMS, FMB, VisDrone) are all publicly available; their processing steps and the stochastic degradation pipeline used to create degraded training pairs are detailed in Appendix A.3. Hyperparameters, optimizer settings, learning rate schedules, and GPU hardware are reported in Section 4. We plan to release the source code and pre-trained models upon acceptance. We believe these resources, combined with the descriptions in the paper, provide sufficient information for independent reproduction of the reported results.

REFERENCES

Bing Cao, Yiming Sun, Pengfei Zhu, and Qinghua Hu. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23555–23564, 2023.

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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11744–11754, 2025.

Jun Chen, Liling Yang, Wei Yu, Wenping Gong, Zhanchuan Cai, and Jiayi Ma. Sdsfusion: A semantic-aware infrared and visible image fusion network for degraded scenes. *IEEE Transactions on Image Processing*, 2025.

Tao Chen, Chuang Wang, Yudong Zhang, Kaijian Xia, and Pengjiang Qian. Dmfusion: Degradation-customized mixture-of-experts with adaptive discrimination for multi-modal image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Zeyang Zhang, Sara Atito, Muhammad Awais, and Josef Kittler. One model for all: Low-level task interaction is a key to task-agnostic image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28102–28112, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 1510–1519. IEEE, 2017.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3489–3497. IEEE, 2021.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements, 2024. URL https://arxiv.org/abs/2410.17725.

Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018.

Jiayang Li, Chengjie Jiang, Junjun Jiang, Pengwei Liang, Jiayi Ma, and Liqiang Nie. Towards unified semantic and controllable image fusion: A diffusion transformer approach. *IEEE transactions on pattern analysis and machine intelligence*, 2025a.

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*, 2025b.

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. *arXiv preprint arXiv:2606.26812*, 2026a.

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*, pp. 104130, 2026b.

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

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation, 2023. URL https://arxiv.org/abs/2308.02097.

Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, and Xin Fan. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019.

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.

Kechen Song, Xiaotong Xue, Hongwei Wen, Yingying Ji, Yunhui Yan, and Qinggang Meng. Misaligned visible-thermal object detection: A drone-based benchmark and baseline. *IEEE Transactions on Intelligent Vehicles*, 9(11):7449–7460, 2024. doi: 10.1109/TIV.2024.3398429.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32:6700–6713, 2022a.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Detfusion: A detection-driven infrared and visible image fusion network. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 4003–4011, 2022b.

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 *International Conference on Machine Learning*, pp. 57571–57586. PMLR, 2025.

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.

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*, 2025.

Linfeng Tang, Yeda Wang, Zhanchuan Cai, Junjun Jiang, and Jiayi Ma. Controlfusion: A controllable image fusion network with language-vision degradation prompts. *Advances in Neural Information Processing Systems*, 38:118304–118327, 2026a.

Linfeng Tang, Yeda Wang, Meiqi Gong, Zizhuo Li, Yuxin Deng, Xunpeng Yi, Chunyu Li, Han Xu, Hao Zhang, and Jiayi Ma. Videofusion: a spatio-temporal collaborative network for multi-modal video fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19559–19569, 2026b.

Wei Tang, Fazhi He, and Yu Liu. Ydtr: Infrared and visible image fusion via y-shape dynamic transformer. *IEEE Transactions on Multimedia*, pp. 1–16, 2022. doi: 10.1109/TMM.2022.3192661.

Qiang Wang, Yi Shen, and Jing Jin. Performance evaluation of image fusion techniques. In Tania Stathaki (ed.), *Image Fusion: Algorithms and Applications*, pp. 469–492. Elsevier, Amsterdam, Netherlands, 2008.

Xue Wang, Zheng Guan, Wenhua Qian, Jinde Cao, Runzhuo Ma, and Cong Bi. A degradation-aware guided fusion network for infrared and visible image. *Information Fusion*, 118:102931, 2025.

Zhishe Wang, Yanlin Chen, Wenyu Shao, Hui Li, and Lei Zhang. Swinfuse: A residual swin transformer fusion network for infrared and visible images. *IEEE Transactions on Instrumentation and Measurement*, 2022. doi: 10.1109/tim.2022.3191664.

Zirui Wang, Jiayi Zhang, Tianwei Guan, Yuhan Zhou, Xingyuan Li, Minjing Dong, and Jinyuan Liu. Efficient rectified flow for image fusion. *Advances in Neural Information Processing Systems*, 38: 22927–22948, 2026.

Han Xu, Xunpeng Yi, Chen Lu, Guangcan Liu, and Jiayi Ma. Urfusion: Unsupervised unified degradation-robust image fusion network. *IEEE Transactions on Image Processing*, 2025.

Han Xu, Yuyang Li, Yunfei Deng, Jiayi Ma, and Guangcan Liu. Deno-if: Unsupervised noisy visible and infrared image fusion method. *Advances in Neural Information Processing Systems*, 38:141649–141671, 2026.

Weimin Xue, Yisha Liu, Fei Wang, Guojian He, and Yan Zhuang. A novel teacher–student framework with degradation model for infrared–visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 73:1–12, 2024.

Luojie Yang, Meng Yu, Lijin Fang, Yi Yang, and Yufeng Yue. Cdmfusion: Rgb-t image fusion based on conditional diffusion models via few denoising steps in open environments. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6314–6320. IEEE, 2025.

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, 2024a.

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, 2024b.

Hao Zhang, Lei Cao, and Jiayi Ma. Text-difuse: An interactive multi-modal image fusion framework based on text-modulated diffusion model. *Advances in Neural Information Processing Systems*, 37:39552–39572, 2024.

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*, 2025.

Hao Zhang, Yanping Zha, Qingwei Zhuang, Zhenfeng Shao, and Jiayi Ma. Robust fusion controller: Degradation-aware image fusion with fine-grained language instructions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12466–12474, 2026a.

Yafei Zhang, Yu Liu, et al. Infrared-assisted single-stage framework for joint restoration and fusion of visible and infrared images under hazy conditions, 2026b.

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.

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*, pp. 5906–5916, 2023a.

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*, 2023b.

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*, pp. 25912–25921, 2024.

Pengfei Zhu, Longyin Wen, Dawei Du, Xiao Bian, Heng Fan, Qinghua Hu, and Haibin Ling. Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7380–7399, 2021. doi: 10.1109/TPAMI.2021.3119563.

A Appendix

A.1 Evaluation Metrics

Let A, B, and F denote the high-quality infrared image, high-quality visible-light image, and restored fused image, respectively. For metric computation, all images are converted to 8-bit grayscale with intensity values in $[0, 255]$; if F is a 3-channel color image obtained by color restoration, its luminance component is used, and A, B, and F below refer to these luminance representations. Six objective metrics are used to evaluate the fusion performance: QNCIE, correlation coefficient (CC), sum of correlations of differences (SCD), mean squared error (MSE), peak signal-to-noise ratio (PSNR), and multi-scale structural similarity (MS-SSIM). These metrics measure global correlation, complementary-information preservation, pixel-level fidelity, and structural similarity from different perspectives.

QNCIE. QNCIE measures the statistical dependency between the fused image and the source images based on their global correlation structure. Before computing the metric, all images are converted to the same single-channel representation and linearly normalized to $[0, 1]$. Let $\mathbf{x}$ and $\mathbf{y}$ denote the vectorized versions of two images, and let $\tilde{\mathbf{x}} = \mathbf{x} - \bar{x}\mathbf{1}$ and $\tilde{\mathbf{y}} = \mathbf{y} - \bar{y}\mathbf{1}$ be their mean-centered versions. Their normalized cross-correlation is defined as

$$r_{XY} = \frac{\tilde{\mathbf{x}}^{\mathsf{T}}\tilde{\mathbf{y}}}{\sqrt{\tilde{\mathbf{x}}^{\mathsf{T}}\tilde{\mathbf{x}}}\sqrt{\tilde{\mathbf{y}}^{\mathsf{T}}\tilde{\mathbf{y}}} + \varepsilon}, \qquad X, Y \in \{A, B, F\}, \tag{14}$$

where ε is a small positive constant used to avoid numerical instability for nearly constant images. The correlation matrix of the two source images and the fused image is then constructed as

$$\mathbf{R} = \begin{bmatrix} 1 & r_{AB} & r_{AF} \\ r_{AB} & 1 & r_{BF} \\ r_{AF} & r_{BF} & 1 \end{bmatrix}. \tag{15}$$

Let s_i, $i = 1, 2, 3$, denote the singular values of $\mathbf{R}$. We first normalize them into a probability distribution:

$$p_i = \frac{s_i}{\sum_{j=1}^{3} s_j}. \tag{16}$$

The correlation information entropy and QNCIE are calculated as

$$H_R = -\sum_{i=1}^{3} p_i \log_2 p_i, \qquad \mathrm{QNCIE} = 1 - \frac{H_R}{\log_2 3}. \tag{17}$$

Here, the convention $0 \log_2 0 = 0$ is adopted. The normalization by $\log_2 3$ constrains the entropy to $[0, 1]$, and consequently $\mathrm{QNCIE} \in [0, 1]$. A higher QNCIE value indicates that the fused image exhibits stronger global statistical dependency with the source images, whereas a lower value indicates a more evenly distributed singular-value spectrum and weaker overall correlation.

Correlation Coefficient. CC measures the linear correlation between the fused image and each source image. For a source image $X \in \{A, B\}$, the correlation with F is

$$r_{XF} = \frac{\sum_i (X_i - \bar{X})(F_i - \bar{F})}{\sqrt{\sum_i (X_i - \bar{X})^2} \sqrt{\sum_i (F_i - \bar{F})^2}}. \tag{18}$$

The final CC value is the average of the two correlations:

$$\mathrm{CC} = \frac{r_{AF} + r_{BF}}{2}. \tag{19}$$

CC reflects the consistency of intensity variations and global structures, but it mainly captures linear similarity.

Sum of Correlations of Differences. SCD evaluates whether the complementary differences between the two source images are preserved. It is defined as

$$\mathrm{SCD} = \mathrm{corr}(F - A,\, B) + \mathrm{corr}(F - B,\, A), \tag{20}$$

where $\mathrm{corr}(\cdot, \cdot)$ denotes the Pearson correlation coefficient. A higher SCD indicates better preservation of inter-modal differences.

Mean Squared Error. After normalizing all images to $[0, 1]$ via $\tilde{X} = X/255$, MSE is computed as

$$\mathrm{MSE} = \frac{1}{2} \left[\frac{1}{MN} \sum_{i=1}^{MN} (\tilde{F}_i - \tilde{A}_i)^2 + \frac{1}{MN} \sum_{i=1}^{MN} (\tilde{F}_i - \tilde{B}_i)^2 \right]. \tag{21}$$

A lower MSE indicates smaller average pixel-level deviations and better intensity fidelity.

Peak Signal-to-Noise Ratio. PSNR is directly derived from MSE:

$$\mathrm{PSNR} = -10 \log_{10}(\mathrm{MSE}). \tag{22}$$

A higher PSNR corresponds to a lower pixel error, but it does not always reflect perceptual quality.

Multi-Scale Structural Similarity. MS-SSIM extends SSIM to multiple scales. For two images X and Y, the multi-scale similarity is defined as

$$\text{MS-SSIM}(X, Y) = [l_M(X, Y)]^{\alpha_M} \prod_{j=1}^{M} [c_j(X, Y)]^{\beta_j} \, [s_j(X, Y)]^{\gamma_j}, \tag{23}$$

where l_M, c_j, and s_j denote luminance, contrast, and structure similarities at scale j, and the weights are typically set to 1. The final metric is

$$\text{MS-SSIM} = \frac{\text{MS-SSIM}(A, F) + \text{MS-SSIM}(B, F)}{2}. \tag{24}$$

MS-SSIM reflects structural and perceptual consistency across scales.

These metrics are complementary. CC and MS-SSIM emphasize structural consistency, SCD evaluates complementary-information preservation, MSE and PSNR measure pixel-level fidelity, and QNCIE describes the global correlation structure. Therefore, the fusion performance should be assessed using multiple metrics rather than a single indicator.

A.2 AdaConv & CondMod

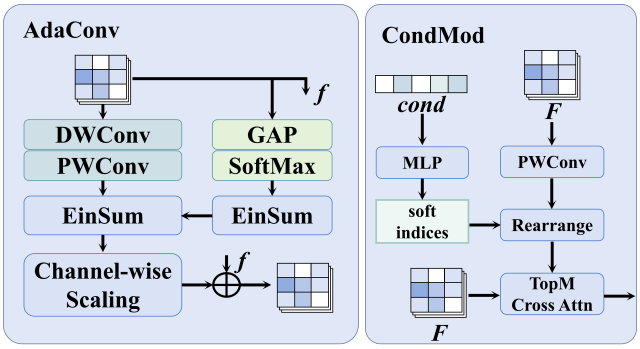

Figure 7: Architecture of the two adaptation modules in Stage I: AdaConv and CondMod.

As is shown in Fig. 7, Stage I employs two complementary adaptation modules, AdaConv and CondMod, which respectively perform input-adaptive local filtering and condition-aware residual cross-attention modulation.

AdaConv. Within each encoder or decoder block, AdaConv performs input-adaptive local filtering: given feature $f \in \mathbb{R}^{B\times C\times H\times W}$, a depthwise convolution and pointwise expansion extract local structure, while a globally pooled style vector

$$a(f) = \mathrm{softmax}\left(W_{\mathrm{style}} \cdot \mathrm{GAP}(f)\right) \tag{25}$$

creates 8 coefficients. These coefficients are linearly projected to the expansion dimension c_{emb} and then used to scale the rows of a learnable matrix S, producing a sample-specific channel-mixing weight via an einsum. The modulated features are projected back and added residually with a near-zero-initialized layer scale γ:

$$\mathrm{AdaConv}(f) = f + \gamma \odot \Phi(f, a(f)). \tag{26}$$

This adapts channel-mixing behavior to the input's own degradation statistics, independent of the explicit wavelet condition c^m, without altering spatial resolution or channel count.

CondMod. While AdaConv adapts to implicit feature statistics, CondMod injects the explicit wavelet condition c^m into the same-scale feature via residual cross-attention:

$$f^m_{k+1} = \tilde{f}^m_{k+1} + \mathcal{C}_k(\tilde{f}^m_{k+1}, c^m). \tag{27}$$

The condition is projected to $2C_k$ channels and used to produce a soft reordering of the feature map; the reordered feature serves as the query in a channel cross-attention against the original feature as key/value. This lets condition-relevant channels be selectively emphasized rather than uniformly rescaled.

A.3 Stochastic Degradation Pipeline

To construct a diverse training set with controllable and reproducible degradation patterns, we implement a stochastic degradation pipeline for visible-infrared image pairs. The pipeline operates on normalized images with pixel values in $[0, 1]$. For each training sample, one of two degradation modes is selected. With probability $p_{\mathrm{sev}} = 0.05$, a predefined severe four-fold degradation is applied. Otherwise, the number of degradation types is sampled from $\{1, 2, 3, 4\}$ with probabilities

$$P(N_{\mathrm{deg}} = 1, 2, 3, 4) = (0.10,\ 0.10,\ 0.30,\ 0.50).$$

Given N_{deg}, N_{deg} distinct degradation types are uniformly sampled without replacement from the degradation pool

$$\mathcal{D} = \{\text{noise}, \text{blur}, \text{stripe}, \text{haze}\}.$$

The selected operators are applied sequentially in the order returned by the random sampling procedure. After each operation, the output is clipped to the valid range $[0, 1]$. Therefore, the training distribution contains both single-degradation samples and compounded-degradation samples with different orders and modality-specific combinations.

Modality-specific Degradation Assignment. The degradation operators are not applied identically to the visible and infrared modalities. Instead, each operator follows a fixed modality affinity: noise is applied only to the visible modality, blur to both the visible and infrared modalities, stripe artifacts to the infrared modality only, and haze to the visible modality only. This asymmetric design is intended to approximate the heterogeneous degradation characteristics of real visible-infrared observations. In particular, additive intensity noise and atmospheric haze are applied to the visible image, while stripe noise is used to model infrared-specific fixed-pattern or column-correlated artifacts.

Noise. For the visible image, Gaussian noise is sampled with probability 0.9 and Poisson noise is sampled with probability 0.1. For Gaussian noise, the standard deviation is sampled from

$$\sigma \sim \mathcal{U}\left(\frac{1}{255}, \frac{60}{255}\right).$$

The noise is clipped to $[0, 1]$ after addition. With probability 0.4, the Gaussian or Poisson noise is generated as grayscale noise shared across the three visible channels; otherwise, channel-wise noise is used. The Poisson noise scale is sampled from

$$\lambda_{\text{P}} \sim \mathcal{U}(0.05, 1.5).$$

Blur. The active blur operator is a disk-shaped defocus blur applied independently to the visible and infrared images. Its radius is fixed to $r = 4$, resulting in a 9×9 kernel. For a pixel location (x, y), the unnormalized kernel is defined as

$$K(x, y) = \max\left(r + 0.5 - \sqrt{x^2 + y^2},\, 0\right),$$

where $x, y \in \{-4, \ldots, 4\}$. The kernel is normalized such that $\sum_{x,y} K(x, y) = 1$ and is applied independently to each channel using reflective padding. This operation models spatial defocus and low-pass degradation shared by the two modalities.

Infrared Stripe. The infrared stripe operator consists of a spatially independent Gaussian component and a column-correlated component. Specifically,

$$I_{\text{stripe}} = I + n_{\text{G}} + n_{\text{col}},$$

where

$$n_{\text{G}} \sim \mathcal{N}(0, \sigma_{\text{G}}^2), \qquad \sigma_{\text{G}} \sim \mathcal{U}\left(\frac{15}{255}, \frac{20}{255}\right),$$

and

$$n_{\text{col}}(x, y) = c(y), \qquad c(y) \sim \mathcal{N}(0, \sigma_{\text{col}}^2), \qquad \sigma_{\text{col}} \sim \mathcal{U}\left(\frac{10}{255}, \frac{15}{255}\right).$$

The same column-correlated noise map is shared by all channels, which approximates the channel-consistent fixed-pattern noise commonly observed in infrared sensors. The resulting image is clipped to $[0, 1]$ before the next degradation is applied.

Haze degradation. Visible-image haze is generated using an inhomogeneous atmospheric transmission model. Let I denote the clean visible image and let

$$A = \text{mean}(I)$$

be the global atmospheric-light value estimated from the input image. For a pixel at position (x, y), the radial distance from the image center is

$$\rho(x, y) = \sqrt{(x - x_c)^2 + (y - y_c)^2},$$

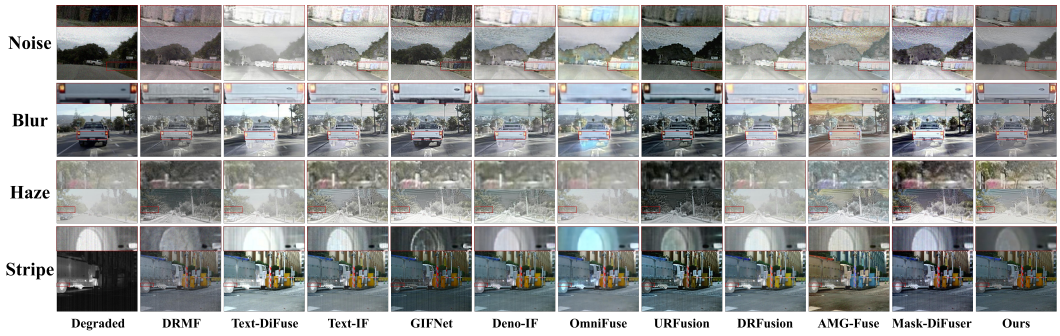

Figure 8: Visual comparison on the EMS dataset for four individual degradation tasks.

where (x_c, y_c) is the image center. The effective scene-depth surrogate is defined as

$$d(x,y) = -c\,\rho(x,y) + \sqrt{\max(H,W)}, \qquad c \sim \mathcal{U}(0.02, 0.06).$$

The transmission map is then computed as

$$t(x,y) = \exp\big(-\beta d(x,y)\big), \qquad \beta = 0.07,$$

and the hazy image is obtained by

$$I_{\text{haze}}(x,y) = I(x,y)t(x,y) + A\big(1 - t(x,y)\big).$$

The output is finally clipped to $[0,1]$. This formulation produces a spatially varying haze strength while preserving a physically interpretable airlight-transmission structure.

Severe composite degradation. With probability $p_{\text{sev}} = 0.05$, all four degradation types are applied in a fixed order:

$$\text{haze} \rightarrow \text{blur} \rightarrow \text{severe noise} \rightarrow \text{infrared stripe}.$$

The haze is applied to the visible image, the disk blur is applied to both modalities, the severe Gaussian/Poisson noise is applied to the visible image, and the stripe noise is applied to the infrared image. Thus, the severe branch corresponds to the multi-hot degradation configuration

$$(\text{noise}, \text{blur}, \text{stripe}, \text{haze}) = (1,1,1,1).$$

This branch accounts for 5% of the training samples and explicitly exposes the model to strongly compounded cross-modal degradation.

A.4 Qualitative and Quantitative Performance on Each Task in the EMS Dataset

We evaluate GranFlow on the EMS dataset across four distinct degradation tasks: noise, stripe artifacts, blur, and haze. For each task, we report both visual comparisons (Fig. 8) and quantitative metrics (Tables 6 and 7). The results demonstrate that our method adapts its correction mechanism to the specific frequency-domain or distributional characteristics of each degradation type, producing consistent improvements over existing approaches.

Noise Degradation. Qualitatively, the noisy input exhibits dense high-frequency sensor noise that obscures fine texture while leaving scene structure largely intact. Among competing methods, Text-DiFuse and OmniFuse fail to suppress noise sufficiently, leaving visible speckle patterns; DRMF and Mask-DiFuser remove noise but overly smooth edges, losing textural details. In contrast, GranFlow achieves a clean result that preserves both object boundaries and high-frequency texture, as the wavelet-conditioned gating module selectively attenuates high-frequency noise while retaining structural information. The scale-aware global conditioning further provides a correction matched to the sample's own displacement scale, ensuring that the flow model does not over-correct the latent representation.

Table 6: Quantitative comparison on the EMS dataset for noise degradation (left) and stripe noise (right).

Method	Noise						Stripe					
	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM
DRMF	0.8701	0.5174	1.8850	0.0709	11.5770	0.5779	0.8714	0.6053	1.8489	0.0490	13.2501	0.6341
Text-DiFuse	0.8447	0.3274	1.2416	0.0896	10.6237	0.3565	0.8708	0.5506	1.1497	0.1372	8.7390	0.6045
Text-IF	0.8616	0.4503	1.5163	0.0629	12.1672	0.5398	0.8804	0.5569	1.6033	0.0803	11.1073	0.6243
GIFNet	0.8675	0.4928	1.5533	0.0620	12.1852	0.4932	0.8689	0.5779	1.5951	0.0491	13.2536	0.6347
Deno-IF	0.8536	0.4443	1.6790	0.0548	12.7508	0.4712	0.8796	0.5924	1.7824	0.0618	12.2493	0.6558
OmniFuse	0.8643	0.4311	1.4956	0.0659	11.9874	0.5603	0.8775	0.5335	1.6649	0.0841	10.8566	0.6325
URFusion	0.8758	0.4691	1.4361	0.1236	9.1309	0.4714	0.8751	0.5306	1.1358	0.0575	12.5071	0.6050
DRFusion	0.8571	0.4452	1.6184	0.0631	12.2025	0.5576	0.8753	0.5586	1.5215	0.0767	11.3272	0.6354
AMG-Fuse	0.8607	0.4576	1.6522	0.0559	12.6886	0.5394	0.8708	0.5475	1.5144	0.0632	12.1644	0.6252
Mask-DiFuser	0.8739	0.5285	1.4646	0.0712	11.5323	0.5392	0.8744	0.6095	1.1089	0.0737	11.6866	0.5919
GranFlow(Ours)	0.8796	0.5150	1.7826	0.0578	12.4797	0.5865	0.8791	0.5741	1.8978	0.0441	13.7219	0.6639

Quantitatively (Table 6, left block), GranFlow attains the best QNCIE and MS-SSIM, indicating better global correlation and multi-scale structural fidelity. It also achieves the second-best SCD and competitive PSNR, with MSE ranking third. This confirms that the wavelet-domain conditioning is highly effective for frequency-selective noise degradation, balancing noise suppression with detail preservation.

Stripe Noise. Stripe corruption introduces periodic vertical high-frequency anomalies that are localized in both space and frequency. Visual inspection shows that DRMF and AMG-Fuse leave residual stripes, while Text-IF and URFusion over-smooth the image, causing loss of texture. GranFlow removes the periodic interference while preserving the underlying scene content, as the gated fusion adaptively downweights the corrupted modality based on the wavelet subband statistics. The global scale embedding also helps the Rectified Flow model to apply a correction magnitude appropriate for stripe degradation.

Quantitative results (Table 6, right block) show that GranFlow achieves the best SCD, MSE, PSNR, and MS-SSIM, with QNCIE ranking third. The higher SCD and PSNR demonstrate that the method effectively eliminates the stripe artifacts while preserving complementary information from both modalities.

Table 7: Quantitative comparison on the EMS dataset for blur (left) and haze (right).

Method	Blur						Haze					
	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM
DRMF	0.8714	0.5220	1.8682	0.0696	11.6680	0.5943	0.8776	0.4208	1.8441	0.0842	10.8927	0.5303
Text-DiFuse	0.8480	0.3358	1.2664	0.0892	10.6567	0.3709	0.8402	0.3350	1.6674	0.0801	11.0584	0.3957
Text-IF	0.8631	0.4573	1.5474	0.0619	12.2534	0.5759	0.8512	0.4221	1.7739	0.0531	12.9032	0.5491
GIFNet	0.8745	0.5131	1.6466	0.0573	12.5521	0.5684	0.8768	0.4459	1.8917	0.0541	12.7596	0.5997
Deno-IF	0.8547	0.4496	1.6693	0.0555	12.6919	0.4796	0.8593	0.3750	1.7214	0.0558	12.6732	0.4266
OmniFuse	0.8646	0.4233	1.4459	0.0683	11.8423	0.5722	0.8616	0.3920	1.7596	0.0700	11.6620	0.5102
URFusion	0.8726	0.4933	1.5455	0.1205	9.2444	0.5144	0.8743	0.4379	1.8372	0.1151	9.4696	0.5587
DRFusion	0.8578	0.4445	1.6164	0.0637	12.1705	0.5833	0.8530	0.4213	1.7989	0.0716	11.5484	0.5296
AMG-Fuse	0.8621	0.4784	1.7496	0.0518	13.0193	0.6017	0.8748	0.4436	1.8732	0.0515	13.0264	0.5774
Mask-DiFuser	0.8736	0.5310	1.3956	0.0773	11.1742	0.5579	0.8734	0.4409	1.5614	0.0616	12.2168	0.5622
GranFlow(Ours)	0.8782	0.5128	1.8725	0.0528	12.8965	0.6028	0.8690	0.4687	1.9168	0.0466	13.4414	0.5919

Blur. Blur represents a loss of high-frequency information, where the fusion model must hallucinate sharp details rather than simply suppress noise. The blurred inputs in Fig. 8 show defocused objects with barely discernible contours. Methods such as Deno-IF and GIFNet produce blurry outputs that fail to recover lost detail, while DRFusion and OmniFuse introduce ringing artifacts. GranFlow generates sharp and naturally contoured results, leveraging the latent Rectified Flow to reconstruct realistic high-frequency content. The conditional summary embedding provides global context, enabling the flow model to generate coherent sharp edges.

Table 7 (left block) quantifies this advantage. GranFlow ranks first in QNCIE, SCD, and MS-SSIM, and second in MSE and PSNR. The SCD value indicates that the fused image preserves the complementary differences between source modalities even after blur removal, which is essential for effective fusion.

Haze. Haze introduces a global, low-frequency distributional shift that primarily affects contrast and color. The hazy inputs in Fig. 8 exhibit a strong veiling effect and faded colors. Competing methods often fail to restore contrast: some (e.g., Text-IF) overcorrect and cause oversaturation, while others (e.g., DRMF) leave residual haze. GranFlow restores natural contrast and color balance, as the scale-aware conditioning measures the latent-space discrepancy and provides an appropriate global correction, avoiding per-pixel artifacts.

Quantitatively (Table 7, right block), GranFlow achieves the best SCD, MSE, PSNR, and CC, with MS-SSIM ranking second. The leading PSNR and MSE indicate strong intensity fidelity, while the high SCD shows that complementary information is well preserved despite the global shift.

A.5 Qualitative and Quantitative Performance on Each Task in the DeMMI-RF Dataset

To further validate the robustness of GranFlow, we evaluate on the DeMMI-RF dataset across six individual tasks: noise at levels $\sigma = 15, 25, 50$, stripe artifacts, defocus blur, and haze. The quantitative results are reported in Tables 8, 9, and 10. Visual comparisons for these tasks are provided in Fig. 4.

Table 8: Quantitative comparison on DeMMI-RF for noise levels $\sigma = 15$ (left) and $\sigma = 25$ (right).

Method	Noise($\sigma = 15$)						Noise($\sigma = 25$)					
	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM
DRMF	0.8790	0.7008	1.7774	0.0368	14.8556	0.6754	0.8791	0.7078	1.7288	0.0310	15.5742	0.6695
Text-DiFuse	0.8726	0.6479	1.2013	0.0825	11.2494	0.6591	0.8769	0.6765	1.3225	0.0627	12.6418	0.6545
Text-IF	0.8777	0.6645	1.6053	0.0561	13.1917	0.6619	0.8790	0.6776	1.5615	0.0370	15.2069	0.6493
GIFNet	0.8719	0.6746	1.4408	0.0423	14.2835	0.6263	0.8714	0.6728	1.3248	0.0382	14.6466	0.5987
Deno-IF	0.8858	0.6211	1.5578	0.0666	13.2896	0.6684	0.8889	0.5737	1.5449	0.0815	13.6318	0.6199
OmniFuse	0.8707	0.6343	1.5466	0.0636	12.6842	0.6683	0.8771	0.6698	1.6081	0.0453	14.4056	0.6699
URFusion	0.8792	0.6849	1.5527	0.0475	13.5954	0.6586	0.8819	0.7093	1.5375	0.0481	13.4522	0.6310
DRFusion	0.8773	0.6332	1.4137	0.0602	12.9586	0.6774	0.8810	0.6715	1.4918	0.0450	14.3005	0.6816
AMG-Fuse	0.8803	0.6550	1.5083	0.0570	13.4174	0.6890	0.8808	0.6825	1.5352	0.0395	14.9768	0.6728
Mask-DiFuser	0.8853	0.7015	0.9714	0.0638	12.2779	0.6280	0.8878	0.7134	0.9778	0.0569	12.7876	0.6201
GranFlow(Ours)	0.8867	0.7052	1.7027	0.0386	14.7620	0.7113	0.8916	0.7307	1.7418	0.0280	16.1108	0.7337

Noise ($\sigma = 15, 25$). Table 8 quantifies the performance. For noise($\sigma = 15$), GranFlow achieves the best QNCIE, CC, and MS-SSIM, and second-best SCD and MSE. For noise($\sigma = 25$), it attains the best scores across all six metrics: QNCIE, CC, SCD, MSE, PSNR, and MS-SSIM. Table 8 shows that GranFlow adapts its correction strength to the noise level, improving as severity increases.

Table 9: Quantitative comparison on DeMMI-RF for heavy noise (left) and stripe artifacts (right).

Method	Noise($\sigma = 50$)						Stripe					
	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM
DRMF	0.8603	0.6071	1.4714	0.0342	15.1849	0.5900	0.8858	0.7309	1.8460	0.0283	16.0312	0.7029
Text-DiFuse	0.8659	0.6179	1.2453	0.0618	12.5497	0.5906	0.8732	0.6653	1.3533	0.0738	11.8929	0.6172
Text-IF	0.8579	0.5747	1.2546	0.0404	14.4390	0.5479	0.8625	0.6114	1.2259	0.0396	14.4424	0.5197
GIFNet	0.8480	0.5350	1.0107	0.0526	13.0657	0.4851	0.8617	0.6301	1.2240	0.0415	14.0940	0.5332
Deno-IF	0.8821	0.4779	1.4031	0.0954	12.6978	0.5351	0.8850	0.6588	1.6528	0.0458	15.1950	0.6536
OmniFuse	0.8670	0.6184	1.5240	0.0440	14.3491	0.5927	0.8731	0.6552	1.6054	0.0443	14.4858	0.6993
URFusion	0.8702	0.6471	1.3948	0.0449	13.7279	0.5481	0.8741	0.6734	1.4163	0.0458	13.6916	0.5563
DRFusion	0.8708	0.6199	1.3777	0.0462	14.0206	0.6301	0.8724	0.6342	1.4183	0.0468	14.0454	0.6138
AMG-Fuse	0.8681	0.6237	1.3915	0.0434	14.3725	0.6140	0.8692	0.6374	1.4491	0.0427	14.5467	0.6004
Mask-DiFuser	0.8813	0.6674	0.9237	0.0555	12.9858	0.5723	0.8769	0.6952	1.1241	0.0499	13.2033	0.5557
GranFlow(Ours)	0.8851	0.6956	1.6730	0.0287	15.9961	0.7089	0.8870	0.7130	1.6166	0.0330	15.4265	0.7074

Noise ($\sigma = 50$). As shown in Table 9 (left block), GranFlow achieves the best performance on all six metrics: QNCIE, CC, SCD, MSE, PSNR, and MS-SSIM. The substantial improvement in PSNR (more than 0.8 dB over the second-best method) indicates that the Rectified Flow efficiently reconstructs the clean latent even from heavily corrupted inputs.

Stripe Noise. Quantitatively (Table 9, right block), GranFlow achieves the best QNCIE and MS-SSIM, and the second-best CC, MSE and PSNR. The leading QNCIE and MS-SSIM indicate that the periodic interference is removed without sacrificing structural consistency across scales.

Table 10: Quantitative comparison on DeMMI-RF for defocus blur (left) and haze (right).

Method	Defocus blur						Haze					
	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM
DRMF	0.8815	0.7071	1.7583	0.0420	14.1091	0.6797	0.8792	0.6553	1.7517	0.0421	14.3222	0.6717
Text-DiFuse	0.8760	0.6688	1.2210	0.0886	10.9012	0.6571	0.8729	0.5931	1.1853	0.0882	11.2791	0.6546
Text-IF	0.8830	0.6811	1.5646	0.0516	13.6108	0.6998	0.8844	0.5992	1.3982	0.0513	13.9181	0.6780
GIFNet	0.8835	0.7066	1.6501	0.0363	15.3921	0.7210	0.8755	0.6275	1.6942	0.0420	14.5972	0.6963
Deno-IF	0.8871	0.6238	1.4967	0.0659	13.6190	0.6634	0.8844	0.5146	1.3968	0.0797	12.6694	0.6183
OmniFuse	0.8731	0.6511	1.5219	0.0639	12.4765	0.6586	0.8682	0.5866	1.5525	0.0703	12.5593	0.6549
URFusion	0.8822	0.6873	1.5488	0.0502	13.4311	0.6928	0.8783	0.6102	1.5057	0.0476	13.7371	0.6809
DRFusion	0.8811	0.6592	1.5328	0.0524	13.7069	0.6933	0.8765	0.5689	1.2638	0.0663	12.7777	0.6562
AMG-Fuse	0.8844	0.6642	1.5287	0.0498	14.0731	0.7105	0.8847	0.5954	1.3432	0.0697	12.8567	0.6876
Mask-DiFuser	0.8865	0.7121	1.0618	0.0613	12.4578	0.6735	0.8864	0.6755	1.0124	0.0603	12.5586	0.6521
GranFlow(Ours)	0.8917	0.7103	1.9261	0.0300	16.1600	0.7436	0.8849	0.6624	1.8640	0.0409	14.9796	0.7197

Defocus blur. Table 10 (left block) shows that GranFlow achieves the best QNCIE, SCD, MSE, PSNR, and MS-SSIM, with CC ranking second. The improvements in PSNR and MS-SSIM are substantial, with PSNR improving by nearly 0.8 dB over the second-best method. This validates the effectiveness of the generative flow in recovering lost detail.

Haze. Quantitative results (Table 10, right block) demonstrate that GranFlow attains the best SCD, MSE, PSNR and MS-SSIM, with QNCIE and CC ranking second. The leading PSNR and MS-SSIM confirm that the method effectively removes the atmospheric veil while preserving structural and textural details, outperforming all baselines on SCD, MSE, PSNR, and MS-SSIM.

A.6 Multimodal Restoration-Fusion under Composite Degradation

To further evaluate the robustness of the proposed method under realistic complex conditions, we conduct quantitative comparisons on two types of composite degradations: a simple combination of haze and blur, and a more severe four-fold combination of haze, blur, noise, and stripe artifacts. The quantitative results on the DVTOD and M3SVD datasets are presented in Tables 11 and 12.

Table 11: Quantitative comparison for compound haze+blur degradation on DVTOD (left) and M3SVD (right).

Method	DVTOD						M3SVD					
	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM
DRMF	0.8774	0.6003	1.7982	0.0341	14.8903	0.6335	0.8718	0.5725	1.8314	0.0505	13.7974	0.6236
Text-DiFuse	0.8559	0.5308	1.6479	0.0644	12.0646	0.6283	0.8606	0.5458	1.6786	0.0831	11.4116	0.6553
Text-IF	0.8595	0.5605	1.6995	0.0388	14.3478	0.6468	0.8600	0.5456	1.7861	0.0515	14.0376	0.6698
GIFNet	0.8856	0.5853	1.8357	0.0327	15.0771	0.6713	0.8801	0.5512	1.8348	0.0431	14.6902	0.6883
Deno-IF	0.8827	0.5841	1.8023	0.0356	14.7192	0.6614	0.8720	0.5656	1.7957	0.0454	14.5427	0.6856
OmniFuse	0.8571	0.5348	1.6689	0.0445	13.7731	0.6107	0.8625	0.5381	1.7784	0.0618	13.3235	0.6520
URFusion	0.8750	0.6104	1.7941	0.0366	14.5961	0.6765	0.8662	0.5777	1.8075	0.0470	14.1497	0.6832
DRFusion	0.8580	0.5369	1.6752	0.0447	13.7529	0.5981	0.8610	0.5307	1.7664	0.0596	13.6246	0.6433
AMG-Fuse	0.8759	0.5685	1.7332	0.0363	14.6464	0.6644	0.8662	0.5272	1.6068	0.0472	14.3054	0.6709
Mask-DiFuser	0.8757	0.6386	1.2583	0.0526	12.9209	0.6516	0.8686	0.5872	1.5158	0.0527	13.8391	0.6851
GranFlow(Ours)	0.8787	0.6326	1.8122	0.0299	15.4409	0.6702	0.8691	0.5851	1.8585	0.0410	15.0464	0.6884

For the **haze&blur** scenario (Table 11), GranFlow achieves the best MSE and PSNR on DVTOD and the best CC, SCD, MSE and PSNR on M3SVD. Although methods like GIFNet show competitive performance on QNCIE, they struggle to maintain pixel-level fidelity under compound corruptions, while GranFlow consistently demonstrates strong structural and intensity fidelity.

For the more challenging **haze&blur&noise&stripe** scenario (Table 12), the advantage of GranFlow becomes even more prominent. On the DVTOD dataset, it ranks first in SCD, MSE, PSNR, and MS-SSIM, while securing the second-best CC and third-best QNCIE. On the M3SVD dataset, GranFlow achieves the best scores across all five metrics including QNCIE, SCD, MSE, PSNR and MS-SSIM, with CC ranking second.

In contrast, many competing methods suffer significant performance degradation when handling the combination of frequency-selective and distributional-shift corruptions. These results indicate that our granularity-matched conditioning mechanisms, specifically the pixel-level wavelet gating for localized degradation and the sample-level displacement-scale signal for distributional shifts, work

Table 12: Quantitative comparison for severe four-fold composite degradation (haze+blur+noise+stripe) on DVTOD (left) and M3SVD (right).

Method	DVTOD						M3SVD					
	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM
DRMF	0.8435	0.4479	1.5100	0.0409	14.0384	0.4538	0.8318	0.3623	1.4489	0.0565	12.7026	0.4252
Text-DiFuse	0.8503	0.4936	1.5903	0.0857	10.8479	0.4456	0.8492	0.4824	1.6150	0.1111	9.7808	0.4170
Text-IF	0.8266	0.3360	1.3606	0.0586	12.5344	0.3699	0.8244	0.3066	1.3727	0.0764	11.6347	0.3521
GIFNet	0.8319	0.3580	1.0886	0.0638	12.0348	0.3013	0.8264	0.3024	1.1639	0.0606	12.6765	0.3048
Deno-IF	0.8679	0.5585	1.7252	0.0356	14.7319	0.5371	0.8543	0.5036	1.7222	0.0428	14.3810	0.5243
OmniFuse	0.8464	0.4707	1.6081	0.0474	13.5387	0.4551	0.8478	0.4640	1.6865	0.0568	13.1143	0.4733
URFusion	0.8485	0.5110	1.6014	0.0425	13.9615	0.3999	0.8395	0.4373	1.5790	0.0488	13.5212	0.3741
DRFusion	0.8444	0.4538	1.5788	0.0494	13.3353	0.4223	0.8422	0.4283	1.6262	0.0606	12.9547	0.4184
AMG-Fuse	0.8285	0.3587	1.3640	0.0478	13.3645	0.3491	0.8261	0.3167	1.2950	0.0571	12.9859	0.3198
Mask-DiFuser	0.8728	0.5823	1.2682	0.0485	13.3694	0.4027	0.8578	0.5268	1.5218	0.0475	13.8098	0.3885
GranFlow(Ours)	0.8632	0.5815	1.7856	0.0323	15.2496	0.5985	0.8732	0.5840	1.9244	0.0380	15.0633	0.6616



Figure 9: Qualitative object detection results on the LLVIP dataset.

synergistically, producing consistent and substantial improvements as the degradation complexity increases.

A.7 Quantitative Comparison with Rectified Flow based method

To verify that the improvement of GranFlow does not merely come from adopting a Rectified Flow (RF) backbone, we compare with the RF-based fusion method RFfusion (Wang et al., 2026) under two representative settings: degradation-free and compound Haze+Blur degradation inputs. Table 13 summarizes the results. Under the degradation-free setting, GranFlow improves all six metrics over RF-Fusion, indicating that the proposed conditioning does not compromise clean-input fusion quality. Under Haze+Blur, the improvement is more pronounced. Since both methods are built on a Rectified Flow formulation, the consistent gains suggest that the advantage of GranFlow is not merely inherited from the RF backbone, but arises from matching the conditioning granularity to degradation characteristics: wavelet-conditioned gating for frequency-selective corruption and scale-aware latent-displacement conditioning for distributional shifts. This comparison complements the main experiments and further supports the effectiveness of GranFlow for degradation-robust infrared-visible fusion.

Table 13: Quantitative Comparison with Rectified Flow based method.

Method	Degradation-free						Haze+Blur					
	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM
RFfusion	0.8748	0.5437	1.9290	0.0533	15.2142	0.6713	0.8688	0.5619	1.7649	0.0467	14.1559	0.6301
GranFlow(Ours)	0.8833	0.5457	1.9561	0.0370	15.2704	0.6731	0.8739	0.6088	1.8353	0.0355	15.2436	0.6793

A.8 Object detection results on the LLVIP dataset

We further evaluate the effectiveness of the restored-fused images on the LLVIP dataset using a fixed YOLOv11 detector. The detector is trained for 100 epochs with an input resolution of 640×640, a batch size of 16, and four data-loading workers. The random seed is fixed to 0 for reproducibil-

ity. Unless otherwise stated, the default YOLOv11 optimization and augmentation strategies are retained. The best checkpoint selected according to the validation performance is used for the final evaluation. As shown in Fig. 9, the comparison shows a challenging case with heavily occluded and overlapping targets. While most competing methods (e.g., DRMF, Text-DiFuse, and URFusion) falsely detect it as a single person or several pedestrians due to severe visual ambiguity, our GranFlow successfully preserves the distinct contours of two overlapping individuals, enabling the detector to correctly identify both of them. This demonstrates that the high-fidelity recovery provided by GranFlow is beneficial for downstream perception tasks in low-light and cluttered scenes.

Quantitative results are summarized in Table 5. Our method ranks first on mAP50, mAP50-95, and recall, and second on precision on LLVIP. The significant gains in mAP50-95 over the second-best method indicate the better robustness of our fused images in handling small, partially obscured objects, further confirming the practical utility of GranFlow in real-world applications.

A.9 Semantic segmentation results on the FMB dataset

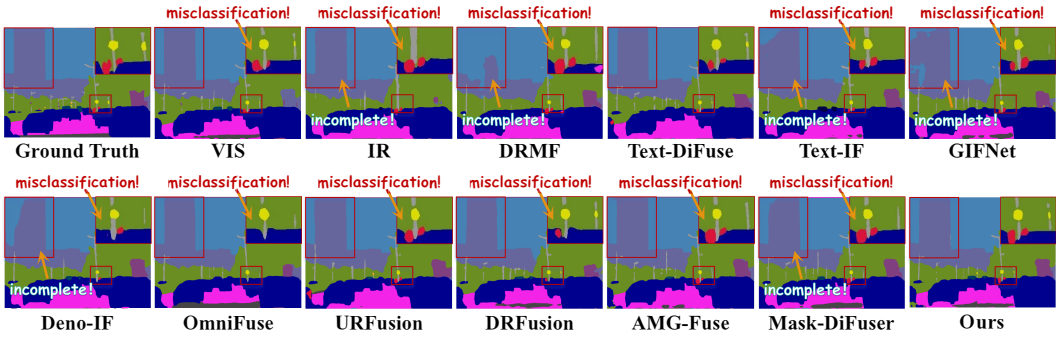

Figure 10: Semantic segmentation performance on the hazy FMB dataset.

Segmentation architecture and evaluation protocol. For each fusion method, we train an independent semantic segmentation model using the corresponding fused RGB images as input. All methods use the same Ultralytics YOLO26-Semantic Small architecture. The model is configured for 14 semantic classes. The input images have a resolution of 800×600.

The FMB data are divided into 1,220 training images and 280 validation images. The same image identifiers, ground-truth masks, class definitions, training and validation splits, initialization, input resolution, optimization settings, and random seed are used for all fusion methods; only the fused input images differ. Thus, the experiment does not perform zero-shot evaluation using a fixed segmentation network. Instead, a separate segmentation model is trained for each fusion method under an identical protocol. Training is performed for up to 100 epochs with batch size selected automatically by the framework, eight data-loading workers, random seed 0, deterministic training enabled, and automatic mixed-precision disabled. The model is evaluated on the validation split after every epoch, and the checkpoint corresponding to the highest validation mIoU is used for reporting.

In Fig. 10, we present visual comparisons across key categories, including building (in light purple), person (in red), pole (in grey), and traffic sign (in yellow). GranFlow generates the most accurate segmentation maps, which align best with the ground truth. Unlike competing methods that suffer from severe misclassification and boundary blurring (e.g., confusing the building and sky regions, or missing the person and pole entirely), GranFlow consistently produces crisp object boundaries and correctly separates overlapping categories, especially in complex scenarios where the person is in front of a pole.

The quantitative evaluation is presented in Table 14. Our method achieves the best mIoU (0.6172), ahead of Mask-DiFuser (0.6034) and GIFNet (0.6019). GranFlow also ranks first or second in multiple semantic classes, including Light, Person, Sky, and Vegetation. This indicates that the spectral and distributional conditioning mechanisms not only enhance the visual quality of the fused images but also effectively preserve and enhance the semantic context required for high-level perception tasks.

Table 14: Comparison of different methods on various semantic categories.

Method	Building	Bus	Car	Light	Motorcycle	Person	Pole	Road	Sidewalk	Sky	Traffic	Traffic Sign	Truck	Vegetation	best mIoU
VI	0.4313	0.5848	0.8098	0.3621	0.5209	0.5474	0.4756	0.2014	0.8623	0.9279	0.7977	0.7447	0.2635	0.8468	0.5983
IR	0.4102	0.0805	0.7653	0.2762	0.4457	0.7091	0.3130	0.1939	0.8608	0.9312	0.8116	0.6175	0.2787	0.8035	0.5355
DRMF	0.3787	0.1320	0.7559	0.1226	0.4191	0.6082	0.3342	0.1706	0.8354	0.8687	0.7017	0.5706	0.2008	0.7664	0.4903
Text-DiFuse	0.4931	0.1486	0.7831	0.3657	0.4470	0.6773	0.4741	0.1924	0.8837	0.9189	0.7718	0.7226	0.3683	0.8442	0.5779
Text-IF	0.4414	0.4994	0.7929	0.2708	0.4209	0.6963	0.4056	0.1802	0.8586	0.9116	0.7743	0.7027	0.3779	0.8323	0.5832
GIFNet	0.4569	0.6399	0.8091	0.3835	0.3525	0.6368	0.4056	0.1579	0.8739	0.9202	0.7931	0.7143	0.4361	0.8471	0.6019
Deno-IF	0.3880	0.5968	0.7965	0.3922	0.4282	0.6986	0.4153	0.1964	0.8606	0.9339	0.8174	0.7369	0.2356	0.8478	0.5961
OmniFuse	0.4630	0.5943	0.8041	0.3290	0.3988	0.6227	0.4389	0.1721	0.8729	0.9122	0.7682	0.7220	0.4157	0.8398	0.5966
URFusion	0.4444	0.5972	0.8081	0.2957	0.5016	0.7048	0.4473	0.1763	0.8709	0.9273	0.7996	0.7349	0.1715	0.8477	0.5950
DRFusion	0.4429	0.1398	0.7886	0.2741	0.4181	0.6544	0.4488	0.2048	0.8694	0.9190	0.7725	0.7426	0.3570	0.8409	0.5623
AMG-Fuse	0.4890	0.1897	0.8171	0.2561	0.5088	0.6741	0.4577	0.2182	0.8846	0.9337	0.8068	0.7286	0.4917	0.8605	0.5941
Mask-DiFuser	0.5388	0.1620	0.8081	0.3828	0.5176	0.6825	0.4668	0.2086	0.8839	0.9322	0.8046	0.7370	0.4608	0.8613	0.6034
GranFlow(Ours)	0.4441	0.6194	0.8101	0.4086	0.5433	0.7097	0.4810	0.1623	0.8846	0.9352	0.8020	0.7183	0.2743	0.8651	0.6172

A.10 Parameter Experiments on Sampling Steps

Table 15: Sampling Step Parameter Experiments on DVTOD.

Step Num	QNCIE	CC	SCD	MSE	PSNR	MS-SSIM	FPS
5	0.8559	0.6033	1.8184	0.0360	14.9558	0.6688	0.52
10	0.8573	0.5890	1.7942	0.0377	14.8390	0.6599	0.28
20	0.8721	0.5915	1.7844	0.0357	14.9842	0.6664	0.14
25	0.8739	0.6088	1.8353	0.0355	15.2436	0.6793	0.11
30	0.8581	0.6067	1.8235	0.0350	15.0402	0.6687	0.09
50	0.8738	0.5945	1.7861	0.0366	14.8636	0.6627	0.06
100	0.8590	0.5991	1.7927	0.0366	14.8672	0.6628	0.03

We evaluate the number of reverse Euler steps in Table 15. The FPS values are measured on 1280×1024 images, matching the full evaluation resolution. Performance improves up to 20 steps and plateaus from 25 to 30 steps. The 25-step setting gives the best overall trade-off, achieving the highest QNCIE, CC, SCD, PSNR, and MS-SSIM at 0.11 FPS. Fewer steps run faster but degrade PSNR and MS-SSIM, while more steps bring no consistent gain but only slow inference. We therefore use 25 steps by default.

A.11 Computational Efficiency and Parameter Experiments

Table 16: Comparison of Computational Efficiency and Parameters

Method	Type	FLOPs(G)	FLOPs/step (G)	Params(M)	FPS	SCD	PSNR
Text-IF	Non-Diff	678.38	678.38	215.12	21.18	1.45	14.09
GIFNet	Non-Diff	103.32	103.32	0.82	80.46	1.42	14.44
Deno-IF	Non-Diff	55.29	55.29	1.43	20.39	1.50	13.50
URFusion	Non-Diff	27.05	27.05	0.24	256.79	1.50	13.59
AMG-Fuse	Non-Diff	704.74	704.74	60.07	5.89	1.46	14.02
OmniFuse	Diff	438.92	2.19	173.34	0.16	1.56	13.39
DRFusion	Diff	737.78	245.93	31.77	15.59	1.42	13.62
DRMF	Diff	4411.64	441.16	170.98	7.74	1.72	14.92
Mask-DiFuser	Diff	11943.27	2388.65	172.33	2.54	1.01	12.68
Text-DiFuse	Diff	95407.33	3816.29	119.46	0.36	1.25	11.67
GranFlow(Ours)	Diff	6038.34	241.53	169.29	0.97	1.77	15.61

Table 16 compares computational cost, parameter size, and fusion quality on the unified dehaze pairs of the DeMMI-RF dataset. To ensure a fair architecture-level comparison, FLOPs and FPS are measured with a fixed 256×256 random-content input; thus they are content-independent and reflect network overhead only. These FPS values are not directly comparable with the 1280×1024 FPS reported above. For diffusion models, total FLOPs depends on the sampling steps, so FLOPs/step is a more architecture-relevant indicator.

Among diffusion-based fusion methods, GranFlow achieves the best SCD (1.77) and PSNR (15.61). Its per-step computation is 241.53G FLOPs, lower than DRFusion (245.93G), DRMF (441.16G),

Mask-DiFuser (2388.65G), and Text-DiFuse (3816.29G). Its parameter size is 169.29MB, comparable to or slightly lower than DRMF, OmniFuse, and Mask-DiFuser. Compared with the heavier high-quality diffusion baselines, GranFlow reduces total FLOPs by about 49.4% relative to Mask-DiFuser and 93.7% relative to Text-DiFuse, while achieving higher SCD/PSNR. Non-diffusion methods are naturally lighter and faster, but their fusion quality is lower; this reflects the inherent trade-off between compact feed-forward fusion and iterative diffusion-based refinement. Overall, GranFlow provides the most favorable quality-efficiency balance among diffusion-based methods: it achieves the highest fusion accuracy with competitive per-step computation and parameter size, and its FPS remains higher than several diffusion baselines such as OmniFuse and Text-DiFuse.