DECISION-CENTRIC FUSION: A UNIFIED DECISION VARIABLE FOR MODALITY ALIGNMENT AND TASK ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Infrared-visible image fusion is challenged by spatially varying modality reliability: visible images provide rich texture but are vulnerable to adverse conditions, whereas infrared images preserve thermal targets but contain limited fine structure. A key limitation of existing pipelines is that cross-modal information is often introduced after modality-specific feature transformations, or is not propagated to subsequent capacity, attention, and fusion decisions. Thus, different fusion modules may operate under different spatial reliability conditions even under end-to-end training. We propose Decision-Centric Fusion, which addresses this through early bidirectional alignment and persistent spatial decision sharing. The Zip-Alignment Module establishes cross-modal communication before feature decisions are finalized, and a unified decision variable provides a shared spatial conditioning signal for the main fusion modules. We further introduce a Boundary-Aware Refiner to characterize the task-dependent trade-off between perceptual enhancement and downstream performance, and a distillation strategy that compresses the model to 0.25M parameters. Experiments on four public datasets show strong texture preservation, competitive structural fidelity and downstream performance, and effective lightweight deployment.

1 INTRODUCTION

Infrared and visible fusion enhances image representation by integrating complementary information from both modalities, which is critical for downstream tasks such as object detection and semantic segmentation. In industrial applications like autonomous driving, infrared-visible fusion empowers algorithms to accurately recognize pedestrians and vehicles under complex illumination conditions, serving as a key factor for system safety and thus attracting sustained attention.

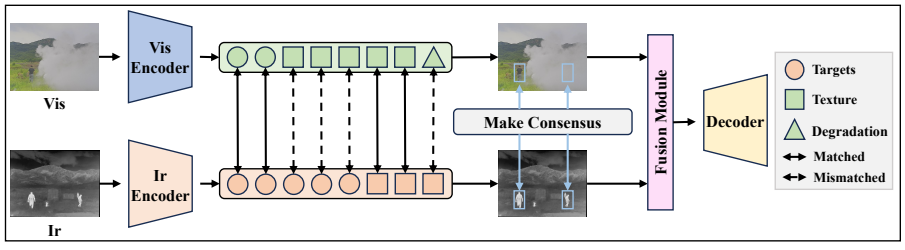

Figure 1: Cross-modal decision lag and shared consensus.Existing methods defer cross-modal interaction to the final fusion stage, leaving stages without a shared modality reliability judgment. DCFusion introduces a shared decision signal that enables visible and infrared modalities to form a consistent consensus across stages.

Current mainstream methods adopt CNNs, Transformers, or generative models as backbones, relying on advanced fusion techniques to directly perform hard fusion on the two modalities.Liu et al. (2025b); Wang et al. (2026b)However, existing methods differ in how early cross-modal information

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.

influences feature decisions. SomeGuan et al. (2026); Sun et al. (2026) defer cross-modal interaction to the final fusion stage, so unimodal features are processed without access to complementary evidence. OthersPark et al. (2025) introduce early interaction but do not carry the interaction result forward to subsequent capacity selection, query construction, and fusion weighting. In both cases, different modules may operate under different spatial reliability conditions, even when the network is trained end-to-end.This introduces our core research question: how can we ensure that cross-modal information is exchanged before feature decisions are finalized, and that the exchanged information persistently conditions subsequent fusion decisions?

Existing approaches broadly fall into two categories to address modality imbalance, yet both lack a shared cross-modal decision signal. One categoryPark et al. (2025); Chen et al. (2026a) employs gating or routing mechanisms to adjust fusion weights based on external conditions such as illumination or semantics. The otherGuan et al. (2026); Zhong et al. (2025) draws on domain adaptation, using dual-rank adapters to capture cross-domain shared knowledge and target-domain specific knowledge. Despite their differences, these methods share a common limitation: cross-modal information either arrives too late to influence feature formation, or is not persistently shared across subsequent decisions. As a result, each stage forms its modality reliability judgment independently, and fusion quality degrades precisely in the regions that need the most attention.

It is precisely this blind spot that motivates our rethinking. We address this at the level of system decision making: a unified decision variable g serves as a shared decision signal, enabling all stages to see where degradation occurs and where fusion should be strengthened, thus reaching a consistent fusion consensus. We call this framework **Decision-Centric Fusion(DCFusion).** The core of our framework lies in three native components. The System Decision Module generates a unified decision variable that serves as the shared decision signal. The Zip-Alignment Module aligns the two modalities like a zipper through four interlocking gears, progressing from coarse semantics to fine textures and from feature capacity to attention sources. The Boundary-Aware Refiner identifies a safe operating region where perceptual quality improves while downstream task performance remains stable.

- We introduce a unified decision variable that provides a shared decision signal for all stages, enabling them to reach a consistent modality reliability judgment. We analyze its gradient paths: G receives gradients from self-adaptive rank decomposition, query replacement, and fusion weighting, whereas independent gates receive only local gradients. This provides cross-module coordination within the fusion network.
- Zip-Alignment Module. We design a Zip-Alignment Module that aligns the two modalities like a zipper through four interlocking gears: encoder attention alignment with infrared high-frequency injection, adaptive rank decomposition, soft query replacement, and decision-guided fusion weighting. Each gear depends on the previous one, forming an interlocking alignment chain.
- Boundary-Aware Refiner. We systematically explore the gain boundary of post-processing refinement and identify a safe operating region where fusion quality improves monotonically while downstream task performance remains stable. This boundary is well-defined and reproducible, providing a practical guideline for deploying refinement in task-sensitive applications.
- Knowledge distillation. We introduce a knowledge distillation strategy that compresses the full pipeline from 1.95M to 0.25M parameters. The distilled model retains 99.7% of the teacher's detection accuracy and achieves comparable fusion quality on all four benchmarks.

Experiments on four public datasets validate the effectiveness of DCFusion, achieving best or second-best results on most metrics while remaining competitive on downstream tasks.

2 Related work

2.1 Deep learning-based Image Fusion

Mainstream infrared-visible fusion methods built on CNNs, Transformers, or generative models have each established mature technical routes. CNN-based approachesMa et al. (2026); Liu et al.

(2025c); Wang et al. (2025); Liu et al. (2025a; 2026); Shi et al. (2026); Wang et al. (2026a); Li et al. (2026c;d) typically use encoder-decoder architectures with hand-crafted or learnable fusion rules in middle layers. Transformer-based methods Wu et al. (2025b); Zhang et al. (2026); Park et al. (2025) exploit self-attention and cross-attention for global cross-modal interaction. Generative models Yang et al. (2025); Guo et al. (2025)employ GANs or diffusion to produce visually plausible results via adversarial training or iterative denoising. These methods either defer cross-modal interaction to the fusion stage or introduce early interaction without carrying it forward to subsequent decisions. In both cases, stages lack a shared modality reliability judgment, leading to inconsistent fusion decisions.

2.2 Methods for Cross-Modal Decision Lag

Spatially varying modality reliability is an inherent challenge of infrared-visible fusion: visible textures degrade under adverse conditions while infrared remains stable. Recent efforts have tackled this bottleneck from different angles. One line adopts gating or routing mechanisms, CtrlFuse Sun et al. (2026) uses mask prompts for dynamic fusion but lacks degradation perception and global decisions. IVIFormerPark et al. (2025) and IGIANetChen et al. (2026a) leverage illumination conditions as gates to guide weights, yet these gates depend on external priors and miss fine-grained intra-modal interaction. Another line builds on domain adaptation, DAFusionGuan et al. (2026) uses dual-rank adapters for shared and target-specific knowledge but neglects cross-modal alignment. FreDFTWu et al. (2025b), AMDANetZhong et al. (2025), and CDTFusionZhao et al. (2026) employ various domain or attention-based strategies to reduce imbalance, yet they defer cross-modal interaction to the fusion stage, leaving stages without a shared fusion consensus and causing inconsistent modality choices.

The above comparison reveals a technical limitation rather than a task-inherent one. Existing gating or adapter designs do not provide a shared spatial condition that the main fusion modules can condition on. Whether interaction happens early or late, each module forms its own modality reliability estimate without a common signal, which is the technical bottleneck we address.

2.3 Comparison with Exiting Methods

The essential difference from existing methodsGuan et al. (2026); Chen et al. (2026a); Park et al. (2025) is twofold. First, they address local optimization conflicts, while we tackle the cross-modal decision lag that prevents stages from forming a shared fusion consensus. Second, instead of deferring interaction to the final fusion stage, we propose a unified decision variable that serves as a shared decision signal for the main fusion modules. Vertically, this signal coordinates encoding, fusion, and post-processing. Horizontally, a chain of alignment gears enables early cross-modal interaction and progressive alignment.

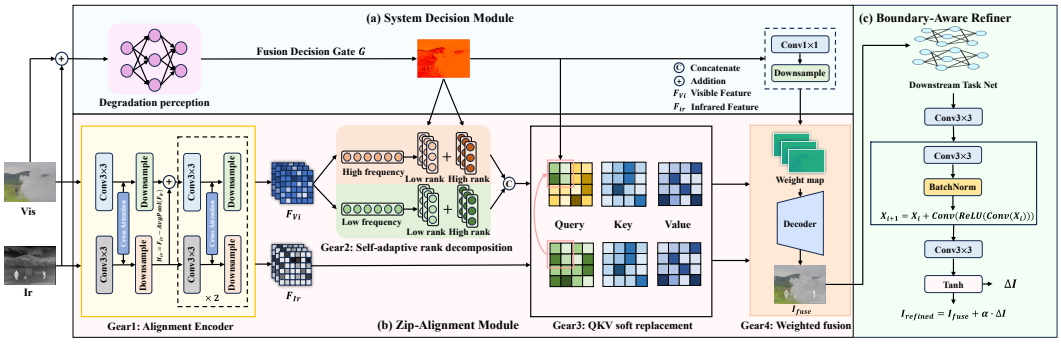

Figure 2: Overview of the Decision-Centric Fusion framework. (a) shows the System Decision Module, which generates the unified decision variable that serves as the global decision signal. (b) illustrates the Zip-Alignment Module, which aligns the two modalities through four interlocking gears: encoder attention alignment with infrared high-frequency injection, adaptive rank decomposition, soft query replacement, and decision-guided fusion weighting. (c) shows the Boundary-Aware Refiner, which identifies a safe operating region for post-processing refinement.

3 METHOD

The architecture of Decision-Centric Fusion is illustrated in Figure 2. Section 3.1 introduces the System Decision Module, which generates the unified decision variable that serves as the shared decision signal across stages. Section 3.2 describes the Zip-Alignment Module, which aligns the two modalities through four interlocking gears. Section 3.3 presents the Boundary-Aware Refiner and its safe operating region. Section 3.4 describes the knowledge distillation strategy.

3.1 SYSTEM DECISION MODULE

We first formulate the proposed decision paradigm. Existing fusion pipelines process features at different feature-processing stages with their own intermediate representations and conditioning signals, without a shared spatial estimate of modality reliability. We introduce a unified decision variable g that serves as the shared decision signal across stages. Eq.(1) is a conceptual formulation, it defines the decision space in which all stages condition on the same variable:

$$\{d_1^*, ..., d_n^*, g^*\} = \arg \min_{d_1,...,d_n,g} \sum_{i=1}^{n} \mathcal{L}_i(d_i, g) + \lambda \cdot \Phi(g) \tag{1}$$

where $\Phi(g)$ regularizes the unified variable. In practice, we adopt the fusion decision gating map G as this decision variable, enabling all stages to form a consistent modality reliability judgment.

We first build a degradation-aware network that perceives the visible degradation distribution:

$$D_{vi} = \sigma(\mathrm{CNN}_{deg}(I_{vi})), \quad D_{vi} \in [0,1]^{H \times W} \tag{2}$$

Based on D_{vi}, we generate the final gating map G:

$$G = \sigma\Big(\mathrm{Conv}_{1\times1}\big(\mathrm{Conv}_{3\times3}(D_{vi})\big)\Big) \tag{3}$$

The unified gating map G is embedded into the gradient paths of three submodules: rank selection $\mathcal{R}$, QKV replacement $\mathcal{Q}$, and fusion weighting $\mathcal{W}$. The total loss gradient w.r.t. G is:

$$\frac{\partial \mathcal{L}_{total}}{\partial G} = \frac{\partial \mathcal{L}_{rank}}{\partial G} + \frac{\partial \mathcal{L}_{qkv}}{\partial G} + \frac{\partial \mathcal{L}_{fuse}}{\partial G} \tag{4}$$

Here $\mathcal{L}_{rank}$, $\mathcal{L}_{qkv}$, and $\mathcal{L}_{fuse}$ are not additional losses; they denote the gradient contributions of the three main fusion submodules to G, all optimized through the main training objective in Eq.(21). This additive gradient structure allows a single update to G to receive signals from all three submodules. In contrast, stage-wise independent gates (G_1, G_2, G_3) yield:

$$\frac{\partial \mathcal{L}_{total}}{\partial G_i} = \frac{\partial \mathcal{L}_i}{\partial G_i}, \quad i = 1, 2, 3 \tag{5}$$

When stage-wise independent gates are optimized only through their local branches, each gate receives only its local gradient. This is the design intuition of our system decision: a shared variable receives cross-module gradient signals that independent gates do not. The unified decision variable G provides this shared conditioning, enabling stages to reach a consistent modality reliability judgment.

3.2 ZIP-ALIGNMENT MODULE

The Zip-Alignment Module aligns the two modalities like a zipper. It consists of four interlocking gears. Gear 1 performs encoder-level attention alignment and infrared high-frequency injection, achieving coarse semantic and fine texture alignment. Gear 2 applies adaptive rank decomposition to align feature capacity according to degradation severity. Gear 3 performs soft query replacement to align attention sources, preventing noisy visible queries from contaminating degraded regions. Gear 4 integrates the dual-path features through decision-guided weight fusion. Each gear depends on the output of the previous one: without Gear 1, Gear 2 lacks aligned features to allocate capacity to; without Gear 2, Gear 3 operates on misaligned features; without Gear 3, Gear 4 cannot determine which modality should dominate at each location. Removing any gear breaks the alignment chain.

Gear 1: Encoder Attention Alignment with Infrared High-Frequency Injection. The infrared and visible encoders share a three-layer CNN downsampling architecture (Figure 2(b)). After each downsampling stage, a lightweight attention module aligns the two modalities:

$$F_{a\leftarrow b} = \text{softmax}\left(\frac{Q_a K_b^\top}{\sqrt{d}}\right) V_b, \quad F_a^{align} = F_a + \gamma \cdot F_{a\leftarrow b} \tag{6}$$

where F denotes the infrared or visible features to be aligned, (a, b) iterates over both directions (vi, ir) and (ir, vi), d is the attention head dimension, and γ is a residual scaling coefficient controlling the bidirectional mapping magnitude. Additionally, to enhance visible stability in degraded regions and achieve deep alignment with infrared, we inject infrared high-frequency texture H_{ir} into visible features F_{vi} after the first downsampling.

$$H_{ir} = F_{ir} - \text{AvgPool}_{3\times3}(F_{ir}) \tag{7}$$

$$F_{vi}^{enh} = F_{vi} + s \cdot H_{ir} \tag{8}$$

where F_{vi}^{enh} is enhanced visible feature and s is the factor that controls injection strength. At this point, we have achieved preliminary semantic alignment of the two modalities at the encoding stage.

Gear 2: Self-Adaptive Rank Decomposition. Visible details in degraded regions are masked by low-frequency noise. We decompose visible features into high-frequency and low-frequency components:

$$F^{lf} = \text{AvgPool}_{7\times7}(F_{vi}), \quad F^{hf} = F_{vi} - F^{lf} \tag{9}$$

where F^{lf} is the low-frequency structural feature obtained via average pooling of F_{vi}, and F^{hf} is the high-frequency residual. Adaptive rank selection is then applied to each component:

$$f_{low}(x) = \text{Conv}_{1\times1}(\text{GELU}(\text{Conv}_{1\times1}(x, r_l)), C) \tag{10}$$

$$f_{high}(x) = \text{Conv}_{1\times1}(\text{GELU}(\text{Conv}_{1\times1}(x, r_h)), C) \tag{11}$$

where x is the input high- or low-frequency feature, and F_{low} and F_{high} are the extracted low-rank r_l and high-rank r_h features. To purify visible information for precise matching with infrared, we introduce the gating map G to perform adaptive rank selection based on degradation degree:

$$F^{lf'} = (1 - G) \cdot f_{low}(F^{lf}) + G \cdot f_{high}(F^{lf}) \tag{12}$$

$$F^{hf'} = (1 - G) \cdot f_{low}(F^{hf}) + G \cdot f_{high}(F^{hf}) \tag{13}$$

Here "low/high-frequency" denotes an AvgPool-based base/detail separation, and "rank" refers to the channel bottleneck width rather than a matrix rank. Both are used as lightweight capacity indicators: low-capacity branches compress shared information, while high-capacity branches preserve faint details.

$$F'_{vi} = F^{lf'} + F^{hf'} \tag{14}$$

After merging both components, we obtain the purified visible feature F'_{vi}, which serves as a fine-grained aligned input for the subsequent cross-modal fusion stage.

Gear 3: Soft Query Replacement. In degraded regions where visible information is severely missing, the attention mechanism can hardly capture valid infrared responses if visible features remain the query source. We thus replace the visible query with the infrared query in such regions:

$$Q_{masked} = (1 - \theta) \cdot Q_{vi} + \theta \cdot Q_{ir} \tag{15}$$

$$\theta = \text{clamp}\left(\frac{G - 0.4}{0.4}, 0, 1\right) \tag{16}$$

where θ is a soft replacement weight derived from the gating map G, adaptively adjusting the injection ratio of infrared queries according to degradation degree. Then, cross-attention yields the visible-based fused feature F_{out}^{vi}, and the infrared counterpart F_{out}^{ir} is computed symmetrically.

$$F_{out}^{vi} = \text{softmax}\left(\frac{Q_{masked} K_{ir}^\top}{\sqrt{d}}\right) V_{ir} \tag{17}$$

Gear 4: Decision-Guided Fusion Weighting. Once the dual-path features are generated, the final step is to integrate them into a single fused feature. Simple additive fusion implicitly assumes equal

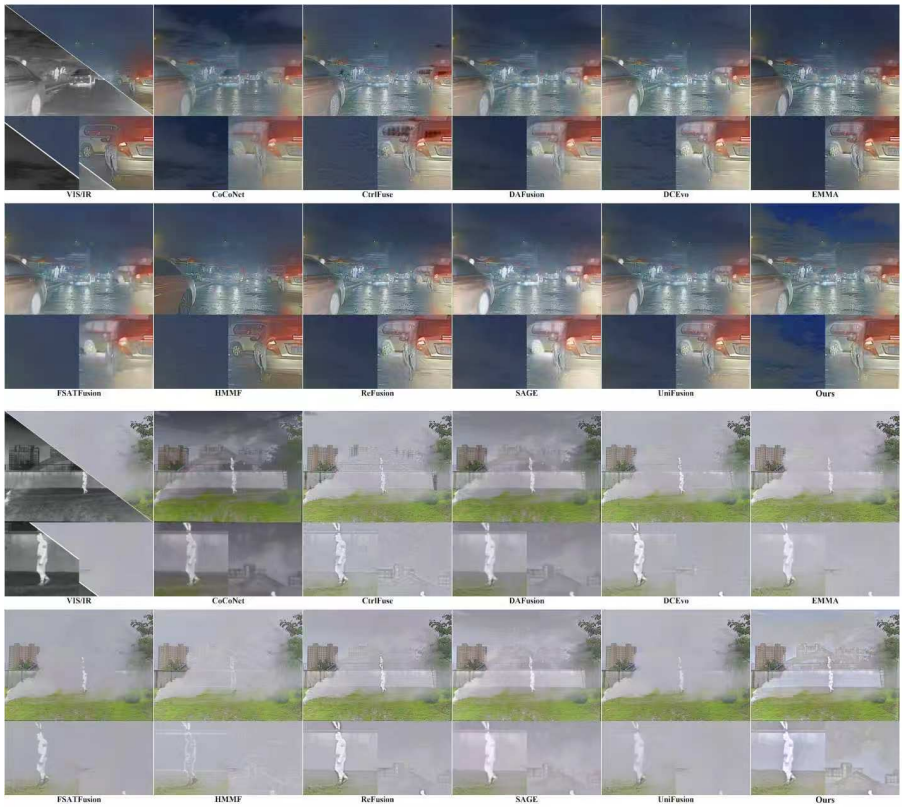

Figure 3: Comparison of fused images generated by our method and other approaches.

contributions from both paths, which essentially remains a hard fusion strategy that ignores the varying reliability of the two modalities across regions. We instead introduce the fusion weight W_{fuse} to make this decision based on global information:

$$W_{fuse} = \sigma(\mathrm{Conv}_{1\times1}(G)) \tag{18}$$

We obtain W_{fuse} by downsampling the gating map G. Since G encodes the shared reliability estimate, it can determine, based on the degradation degree across regions, whether visible or infrared should dominate at each location. The final fused feature is then:

$$F_{fuse} = W_{fuse} \odot F_{out}^{vi} + (1 - W_{fuse}) \odot F_{out}^{ir} \tag{19}$$

At this point, the image fusion stage is completed.

3.3 Boundary-Aware Refiner

The fused image achieves satisfactory visual quality, but the ultimate value of fusion extends beyond appearance to downstream tasks. We design a Boundary-Aware Refiner as an independent exploratory module that fine-tunes the converged fused image. The refiner adopts a residual network (Figure 2(c)), taking the fused image I_{fuse} and infrared features I_{ir} and G as input:

$$I_{refined} = I_{fuse} + \alpha \cdot \Delta I \tag{20}$$

where ΔI is the predicted residual and alpha is a scaling coefficient controlling the refinement magnitude. To preserve the basic structure of the fused image, the gating map G serves as auxiliary guidance to constrain the spatial distribution of refinement strength. The refined output is denoted as $I_{refined}$.

3.4 Knowledge Distillation

To facilitate deployment in real-world scenarios such as autonomous driving and surveillance, we further introduce a knowledge distillation strategy that compresses the full pipeline from 1.95M to

Table 1: Comparison on four datasets (PSNR, EN, SF, AG, SCD). The full model uses $\alpha = 0.10$ at the boundary of the safe operating region. Best results in **bold**, second best underlined.

	M3FD					FMB					MSRS					RoadScene				
Method	PSNR	EN	SF	AG	SCD	PSNR	EN	SF	AG	SCD	PSNR	EN	SF	AG	SCD	PSNR	EN	SF	AG	SCD
ReFusion	16.13	6.91	16.00	6.29	1.60	17.36	6.81	15.17	5.42	1.52	20.39	6.62	**11.63**	**4.39**	1.67	15.67	7.18	6.98	2.97	**1.67**
CoCoNet	15.94	6.67	8.32	3.21	1.50	17.30	6.50	7.67	2.61	1.39	17.80	5.37	7.34	2.53	1.24	15.46	6.94	10.72	4.84	1.30
EMMA	16.21	6.92	15.87	6.51	1.51	17.77	6.69	14.99	5.54	1.51	20.44	6.74	11.55	4.38	1.63	14.69	7.22	7.13	3.20	1.65
SAGE	14.82	6.89	13.37	5.26	**1.71**	15.75	**6.84**	12.64	4.53	1.53	19.57	6.02	10.42	3.70	1.42	14.03	7.06	10.53	4.72	1.50
DCEvo	15.90	6.90	14.84	5.71	1.52	17.19	6.69	13.76	4.83	1.50	20.15	6.64	11.46	4.32	1.67	15.72	7.21	13.01	5.84	1.46
FSATFusion	15.83	6.70	11.93	4.50	1.31	17.28	6.52	11.88	4.17	1.29	20.11	6.60	11.38	4.19	1.54	15.56	6.92	5.69	2.16	1.21
UniFusion	16.31	6.88	14.75	5.81	1.46	18.01	6.65	14.06	4.97	1.45	16.38	6.68	11.16	4.14	1.71	15.65	6.98	4.76	2.24	1.39
CtrlFuse	16.28	6.83	14.90	6.05	1.47	18.00	6.62	13.75	5.08	1.39	20.24	6.43	10.49	4.11	1.49	15.38	7.12	10.66	4.90	1.37
HMMF	16.17	6.64	13.66	5.47	0.91	17.98	6.41	13.46	4.80	0.65	12.91	6.06	7.91	2.84	0.89	12.35	6.91	11.27	4.75	0.59
DAFusion	15.67	6.74	10.41	4.22	**1.71**	16.97	6.68	9.92	3.60	1.56	18.74	5.81	8.24	3.03	1.31	15.52	7.12	7.17	3.26	1.62
Ours	16.35	**6.98**	**19.49**	**7.90**	1.62	**18.03**	6.70	**15.39**	**5.58**	**1.59**	**20.88**	**6.76**	11.09	4.18	**1.79**	**16.20**	**7.28**	**14.52**	**6.92**	1.33
Ours (tiny)	**16.64**	6.92	19.09	7.71	1.62	17.99	6.65	15.11	5.44	1.57	20.58	6.69	10.74	3.98	1.68	16.18	7.27	13.18	6.20	1.35

0.25M parameters. The distilled model retains 99.7% of the teacher's detection mAP50 and achieves comparable or better PSNR and SSIM on all four benchmarks, while the compression ratio reaches 7.7 times and inference speed improves from 21.4 FPS to 35.5 FPS. This demonstrates that the proposed framework can be deployed efficiently with only a small performance drop.

4 EXPERIMENTS

4.1 DATASETS & SETTINGS

We conduct experiments on an RTX 5070Ti (16 GB) across four public benchmarks: M3FD Liu et al. (2022), FMBLiu et al. (2023), MSRSTang et al. (2022), and RoadSceneXu et al. (2022). Dataset splits follow standard protocols or official recommendations from the original sources. The optimizer is AdamW with an initial learning rate of 1×10^{-4}, decayed via cosine annealing to 1×10^{-6}, batch size of 8, and 100 training epochs. The refiner is trained for 150 additional epochs with the main network frozen, while the downstream model provides only inference-level supervision. The total framework has 1.95M parameters, speed 46.81ms with main framework 1.46M and refiner 0.49M.

4.2 MAIN RESULTS

We compare with ten state-of-the-art methods: ReFusionBai et al. (2025), CoCoNetLiu et al. (2024), EMMAZhao et al. (2024), SAGEWu et al. (2025a), DCEvoLiu et al. (2025c), FSATFusionZhang et al. (2026), UniFusionLi et al. (2026b), CtrlFuseSun et al. (2026), HMMFXiao et al. (2026), and DAFusionGuan et al. (2026). This set covers recent advances, well-recognized strong baselines, and methods specifically designed for modality imbalance.

M3FD & FMB Datasets On M3FD, which contains challenging night and foggy scenes, DCFusion achieves the highest SF and AG, outperforming the second-best by 21.8%, and 21.4%, respectively. It also ranks top on EN and PSNR, and is competitive on SCD. The substantial margins on SF and AG indicate richer texture retention, exactly the information most prone to loss in degraded visible images. On FMB, where degradation is milder, performance gaps shrink but DCFusion still leads on PSNR, SF, and AG, with SCD also among the best. This suggests that the advantage of our framework mainly lies in its adaptability to degradation.

MSRS & RoadScene Datasets On MSRS, DCFusion achieves the highest PSNR and SCD, with SCD exceeding the second-best by 4.7%, confirming effective complementary information integration under varying illumination. On RoadScene, EN, SF, and AG all show clear leads, with SF 11.6% and AG 18.5% above the runner-up, validating our method's detail recovery capability in real driving conditions.

Across all four datasets, DCFusion consistently excels on texture-related metrics while remaining competitive on structural fidelity. This pattern substantiates our core claim that addressing cross-

modal decision lag requires recovering lost visible textures in degraded scenes, which inherently involves a trade-off with pixel-level fidelity.

4.3 Downstream Tasks Performance

We evaluate detection on M3FD using YOLOv8nVarghese & M. (2024) trained on the fused images of each method, and segmentation on FMB using SegFormer-B3Xie et al. (2021) trained on the fused images of each method, following the same training protocol for all compared methods.

Table 2: Downstream task performance: segmentation on FMB and detection on M3FD. The full model uses $\alpha = 0.10$ at the boundary of the safe operating region. Best results in **bold**, second best underlined.

Method	**Segmentation (FMB)**		**Detection (M3FD)**	
	mIoU	PA	mAP50	mAP50-95
ReFusion	0.351	<u>0.832</u>	0.719	0.459
CoCoNet	0.272	0.772	0.610	0.401
EMMA	0.341	0.819	0.704	0.466
SAGE	0.464	0.819	0.717	0.476
DCEvo	0.201	0.513	0.715	0.473
FSATFusion	0.462	0.816	0.658	0.424
UniFusion	0.478	0.825	0.713	0.467
CtrlFuse	0.436	0.793	0.706	0.463
HMMF	0.423	0.775	0.699	0.463
DAFusion	0.455	0.814	0.721	**0.490**
Ours	**0.507**	**0.847**	**0.724**	<u>0.478</u>
Ours (tiny)	<u>0.503</u>	0.841	<u>0.723</u>	0.475

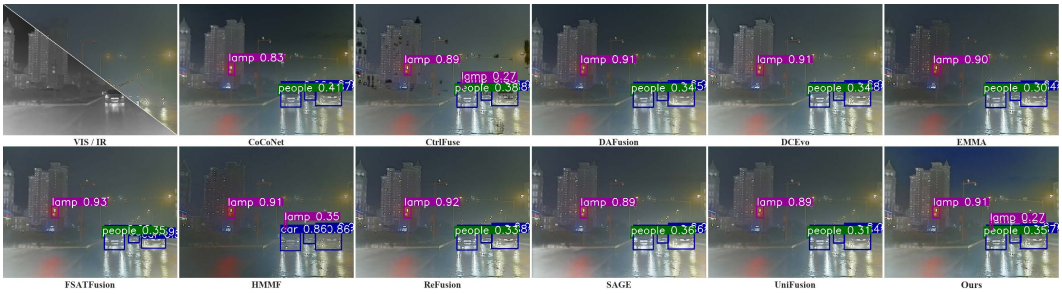

Figure 4: Performance comparison on detection tasks between our fused images and other methods.

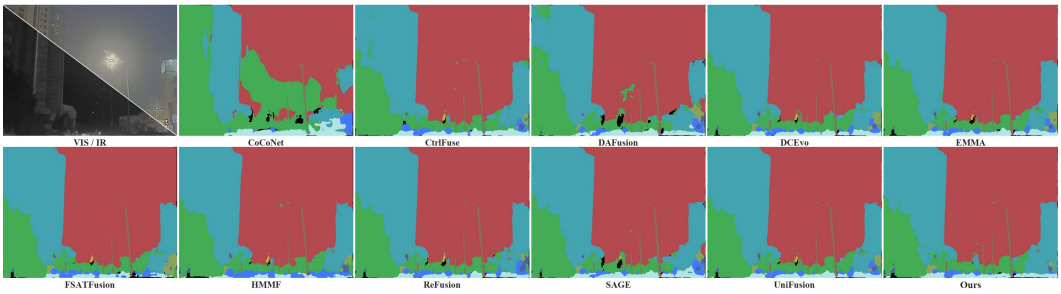

Figure 5: Performance comparison on segmentation tasks between our fused images and other methods.

As reported in Table 2, DCFusion achieves the highest mIoU of 0.507 and PA of 0.847 on FMB, outperforming all competing methods. On M3FD, it attains an mAP50 of 0.724 and mAP50-95

of 0.478. These results demonstrate that the perceptual quality improvements delivered by our framework translate into competitive downstream task performance, confirming the practical value of sharing a decision signal across fusion stages.

4.4 Ablation Study

Table 3: Ablation study on M3FD dataset. Best results are in **bold**, second best underlined.

Ablation	PSNR	EN	SF	AG	SCD	mAP50
Full	16.35	**6.98**	**19.49**	**7.90**	**1.62**	**0.724**
Boundary-Aware Refiner						
w/o Refiner	17.26	6.79	17.13	6.68	1.44	0.716
Zip-Alignment Module						
w/o Align + IR_HF	**17.45**	6.76	16.37	6.28	1.46	0.705
w/o Query soft replace	17.00	6.76	17.01	6.63	1.44	0.713
w/o Rank Decomp	17.05	6.75	16.90	6.59	1.46	0.703
w/o Fuse Weight	17.09	6.80	16.89	6.60	1.45	0.704
System Decision Module						
w/o Gate all	16.75	6.79	16.86	6.57	1.43	0.700
only Gate $\rightarrow$ Rank	16.88	6.80	16.93	6.60	1.45	0.712
only Gate $\rightarrow$ QKV	17.05	6.80	17.03	6.65	1.44	0.711
only Gate $\rightarrow$ Decoder	17.09	6.80	17.00	6.63	1.45	0.702
only Gate $\rightarrow$ Refiner	16.83	6.89	17.30	6.73	1.60	0.715
Independent Gate	16.92	6.77	17.06	6.66	1.42	0.704

We conduct detailed ablation studies to quantify each component. Removing the Boundary-Aware Refiner reduces EN from 6.98 to 6.79 and SCD from 1.62 to 1.44, while mAP50 decreases only slightly from 0.724 to 0.716. This indicates that the refiner mainly improves perceptual quality and its impact on detection remains limited. Within the Zip-Alignment Module, removing encoder alignment and infrared high-frequency injection causes the largest drop, with SF falling from 19.49 to 16.37 and mAP50 from 0.724 to 0.705. Removing adaptive rank decomposition, soft query replacement, or decision-guided fusion weighting reduces mAP50 to 0.703, 0.713, and 0.704 respectively. These results confirm that the interlocking gears jointly reduce cross-modal decision lag and that each gear contributes to the final performance.

For the System Decision Module, removing the gate entirely yields the lowest mAP50 (0.700). Replacing the unified gate with independent gates that receive only local gradients at each fusion stage gives 0.704, while applying the gate to only one stage yields intermediate results (0.702–0.715). Applying the gate only to the refiner gives 0.715, close to w/o Refiner (0.716), indicating that the refiner alone does not account for the full gain. The full model achieves the best mAP50 (0.724), showing that sharing the decision variable across the main fusion modules provides a common spatial condition and improves fusion decisions.

5 Conclusion

In this research, we propose Decision-Centric Fusion to address cross-modal decision lag in infrared-visible fusion. We identify a technical limitation that the main fusion modules lack a shared spatial conditioning signal, which a unified decision variable provides. The framework comprises three native components: the System Decision Module, the Zip-Alignment Module, and the Boundary-Aware Refiner. On four public datasets, it achieves leading texture metrics while remaining competitive on structural fidelity and downstream tasks. The refiner identifies a safe operating region for post-processing, and knowledge distillation compresses the pipeline to 0.25M parameters for lightweight deployment.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for tasks with required disclosure: generating experimental code and analyzing results. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, or designing experimental methodology, and these required disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools for tasks with recommended disclosure: editing the manuscript to improve readability and expression. We have reviewed all AI-assisted work. Specifically, all LLM-generated code was verified and tested for correctness by the authors, and all LLM-generated analysis results were checked against the experimental data. All LLM-edited text was reviewed for factual accuracy and clarity. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

See the ICLR 2027 AI Policy for Authors for more details. This statement should not be more than 1 page.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

This work adheres to the ICLR Code of Ethics. Our study focuses on infrared and visible image fusion, a general-purpose low-level vision task. We use four publicly available benchmark datasets (M3FD, FMB, MSRS, and RoadScene). No synthetic datasets were created. The study does not involve human subjects, sensitive personal data, privacy concerns, or applications that could lead to discrimination or bias. We are committed to ensuring fairness and transparency in both experimentation and reporting.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

To ensure reproducibility, we provide the following resources. The main paper describes the proposed Decision-Centric Fusion framework, including the System Decision Module, Zip-Alignment Module, and Boundary-Aware Refiner, with detailed formulations in Section 3. Implementation details, including dataset splits, training configurations, loss functions, and hyperparameter settings, are provided in Appendix A.2. Anonymous source code, pre-trained models, and test samples are provided as supplementary material. All datasets used (M3FD, FMB, MSRS, and RoadScene) are publicly available. These resources allow the results to be reproduced.

REFERENCES

Haowen Bai, Zixiang Zhao, Jiangshe Zhang, Yichen Wu, Lilun Deng, Yukun Cui, Baisong Jiang, and Shuang Xu. Refusion: Learning image fusion from reconstruction with learnable loss via meta-learning. *Int. J. Comput. Vis.*, 133(5):2547–2567, 2025. doi: 10.1007/S11263-024-02256-8. URL https://doi.org/10.1007/s11263-024-02256-8.

Xiangqi Chen, Dawei Zhang, Li Zhao, Chengzhuan Yang, Zhongyu Chen, Jungang Lou, Zhonglong Zheng, Sang-Woon Jeon, and Hua Wang. Igianet: Illumination guided implicit alignment network for infrared-visible UAV detection. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 3056–3064. AAAI Press, 2026a. doi: 10.1609/AAAI.V40I4.37298. URL https://doi.org/10.1609/aaai.v40i4.37298.

Zhuo Chen, Liangli Ma, Yu Zhang, Feng Qian, Xian Zhang, and Yuxiu Liu. Tcdfusion: Task-guided cross-dimensional feature interaction via dynamic learning for infrared and visible image

fusion. *Optics & Laser Technology*, 2026b. URL https://api.semanticscholar.org/CorpusID:289590886.

Tianwei Guan, Hao Wei, Yuhang Zhou, Jun Ma, Zechen Xu, Zhiying Jiang, Jinyuan Liu, and Xingyuan Li. Domain adaptation guided infrared and visible image fusion. In *AAAI Conference on Artificial Intelligence*, 2026. URL https://api.semanticscholar.org/CorpusID:286661358.

Lin Guo, Xiaoqing Luo, Wei Xie, Zhancheng Zhang, Hui Li, Rui Wang, Zhenhua Feng, and Xiaoning Song. Revisiting generative infrared and visible image fusion based on human cognitive laws. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/8a94ff6f922d995d7d3f4ebf4143e442-Abstract-Conference.html.

Tao Jiang, Hongyang Zhao, Yun Liu, Xingdong Li, and Jing Jin. Tmfusion: A task-oriented meta-learning framework for infrared and visible image fusion. *Infrared Physics & Technology*, 2026. URL https://api.semanticscholar.org/CorpusID:286822449.

Daoheng Li, Mengkai Yan, Jun Li, Jian Yang, and Jianjun Qian. Tdhf: Task-driven hierarchical image fusion under low-light conditions. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(4):4064–4077, 2026a. doi: 10.1109/TCSVT.2025.3635231.

Xingyuan Li, Songcheng Du, Yang Zou, HaoYuan Xu, Zhiying Jiang, and Jinyuan Liu. Unifusion: A unified image fusion framework with robust representation and source-aware preservation. *CoRR*, abs/2603.14214, 2026b. doi: 10.48550/ARXIV.2603.14214. URL https://doi.org/10.48550/arXiv.2603.14214.

Yongsong Li, Yuezhen Jing, Zhengzhou Li, and Abubakar Siddique. Infrared and visible image fusion based on one-dimensional guided filtering and cross-modal weight perception. *Digit. Signal Process.*, 168:105593, 2026c. doi: 10.1016/J.DSP.2025.105593. URL https://doi.org/10.1016/j.dsp.2025.105593.

Yulong Li, Zheng Guan, Xue Wang, and Hao Dai. Mwefusion: Multi-scale wavelet-based feature enhancement network for infrared and visible image fusion. *Digit. Signal Process.*, 176:105991, 2026d. doi: 10.1016/J.DSP.2026.105991. URL https://doi.org/10.1016/j.dsp.2026.105991.

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 (CVPR)*, 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. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8081–8090, 2023. doi: 10.1109/ICCV51070.2023.00745.

Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *Int. J. Comput. Vis.*, 132(5):1748–1775, 2024. doi: 10.1007/S11263-023-01952-1. URL https://doi.org/10.1007/s11263-023-01952-1.

Jinyuan Liu, Xingyuan Li, Zirui Wang, Zhiying Jiang, Wei Zhong, Wei Fan, and Bin Xu. Promptfusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE CAA J. Autom. Sinica*, 12(3):502–515, 2025a. doi: 10.1109/JAS.2024.124878. URL https://doi.org/10.1109/JAS.2024.124878.

Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, Xin Fan, and Risheng Liu. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(4):2349–2369, 2025b. doi: 10.1109/TPAMI.2024.3521416. URL https://doi.org/10.1109/TPAMI.2024.3521416.

Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 2226–2235. Computer Vision Foundation / IEEE, 2025c. doi: 10.1109/CVPR52734.2025.00213. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liu_DCEvo_Discriminative_Cross-Dimensional_Evolutionary_Learning_for_Infrared_and_Visible_Image_CVPR_2025_paper.html.

Xiaowen Liu, Hongtao Huo, Haozhe Cao, Xu Dong, and Jing Li. An anisotropic low-rank self-attention mechanism for infrared and visible image fusion. *Eng. Appl. Artif. Intell.*, 176:114754, 2026. doi: 10.1016/J.ENGAPPAI.2026.114754. URL https://doi.org/10.1016/j.engappai.2026.114754.

Linli Ma, Suzhen Lin, Jianchao Zeng, Yanbo Wang, and Zanxia Jin. Clafusion: Misaligned infrared and visible image fusion based on contrastive learning and collaborative attention. *Comput. Vis. Image Underst.*, 263:104574, 2026. doi: 10.1016/J.CVIU.2025.104574. URL https://doi.org/10.1016/j.cviu.2025.104574.

Incheol Park, Youngwan Jin, Yagiz Nalçakan, Hyeongjin Ju, Sanghyeop Yeo, and Shiho Kim. Iviformer: Illumination-aware infrared-visible image fusion via adaptive domain-switching cross attention. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025 - Workshops, Honolulu, HI, USA, October 19-20, 2025*, pp. 2241–2250. IEEE, 2025. doi: 10.1109/ICCVW69036.2025.00234. URL https://doi.org/10.1109/ICCVW69036.2025.00234.

Liuyan Shi, Rencan Nie, Jinde Cao, Jiang Zuo, and Xiaoli Li. A hierarchical semantic collaboration-based network for infrared and visible image fusion. *Eng. Appl. Artif. Intell.*, 168:114025, 2026. doi: 10.1016/J.ENGAPPAI.2026.114025. URL https://doi.org/10.1016/j.engappai.2026.114025.

Yiming Sun, Yuan Ruan, Qinghua Hu, and Pengfei Zhu. Ctrlfuse: Mask-prompt guided controllable infrared and visible image fusion. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 9260–9268. AAAI Press, 2026. doi: 10.1609/AAAI.V40I11.37884. URL https://doi.org/10.1609/aaai.v40i11.37884.

Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Inf. Fusion*, 83-84:79–92, 2022. doi: 10.1016/J.INFFUS.2022.03.007. URL https://doi.org/10.1016/j.inffus.2022.03.007.

Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pp. 1–6, 2024. doi: 10.1109/ADICS58448.2024.10533619.

Hongli Wang, Wenhua Qian, Xue Wang, Chunlan Zhan, Cong Bi, and Shuang Luo. Frfusion: Flare removal for nighttime infrared and visible image fusion. *Digit. Signal Process.*, 168:105586, 2026a. doi: 10.1016/J.DSP.2025.105586. URL https://doi.org/10.1016/j.dsp.2025.105586.

Rongchao Wang, Zhaofa Zhou, Shuhui Li, and Zhili Zhang. Advances and challenges in infrared-visible image fusion: a comprehensive review of techniques and applications. *Artif. Intell. Rev.*, 59(1):18, 2026b. doi: 10.1007/S10462-025-11426-0. URL https://doi.org/10.1007/s10462-025-11426-0.

Yuhao Wang, Lingjuan Miao, Zhiqiang Zhou, Yajun Qiao, and Lei Zhang. Faster and better: Reinforced collaborative distillation and self-learning for infrared-visible image fusion. *CoRR*, abs/2509.02424, 2025. doi: 10.48550/ARXIV.2509.02424. URL https://doi.org/10.48550/arXiv.2509.02424.

Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every SAM drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 17882–17891. Computer Vision Foundation / IEEE, 2025a. doi: 10.1109/CVPR52734.2025.01666. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wu_Every_SAM_Drop_Counts_Embracing_Semantic_Priors_for_Multi-Modality_Image_CVPR_2025_paper.html.

Wencong Wu, Xiuwei Zhang, Hanlin Yin, Shun Dai, Hongxi Zhang, and Yanning Zhang. Fredft: Frequency domain fusion transformer for visible-infrared object detection. *CoRR*, abs/2511.10046, 2025b. doi: 10.48550/ARXIV.2511.10046. URL https://doi.org/10.48550/arXiv.2511.10046.

Yi Xiao, Jia Wang, Zhu Liu, Di Wang, Jinyuan Liu, and Risheng Liu. A hybrid space model for misaligned multi-modality image fusion. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 11005–11013. AAAI Press, 2026. doi: 10.1609/AAAI.V40I13.38078. URL https://doi.org/10.1609/aaai.v40i13.38078.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1): 502–518, 2022. doi: 10.1109/TPAMI.2020.3012548.

Zejun Yang, Huawei Wei, and Zhisheng Wang. Realportrait: Realistic portrait animation with diffusion transformers. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 9345–9353. AAAI Press, 2025. doi: 10.1609/AAAI.V39I9.33012. URL https://doi.org/10.1609/aaai.v39i9.33012.

Tianpei Zhang, Jufeng Zhao, Yiming Zhu, Guangmang Cui, and Yuhan Lyu. Fsatfusion: Frequency-spatial attention transformer for infrared and visible image fusion. *Comput. Vis. Image Underst.*, 263:104600, 2026. doi: 10.1016/J.CVIU.2025.104600. URL https://doi.org/10.1016/j.cviu.2025.104600.

Wenda Zhao, Wenbo Wang, Haipeng Wang, You He, and Huchuan Lu. Cdtfusion: Crossing domain and task for infrared and visible image fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 48 (2):1456–1469, 2026. doi: 10.1109/TPAMI.2025.3614704. URL https://doi.org/10.1109/TPAMI.2025.3614704.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 25912–25921. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02448. URL https://doi.org/10.1109/CVPR52733.2024.02448.

Haifeng Zhong, Fan Tang, Zhuo Chen, Hyung Jin Chang, and Yixing Gao. Amdanet: Attention-driven multi-perspective discrepancy alignment for rgb-infrared image fusion and segmentation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 10645–10655. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00991. URL https://doi.org/10.1109/ICCV51701.2025.00991.

A Appendix

A.1 Exploration of Refiner Boundary Benefits

The refiner serves as a diagnostic probe rather than a performance booster. For refinement strength α, EN, SF, SCD, and AG increase within $\alpha \in [0, 0.25]$, while SSIM remains largely stable, indicating texture enhancement without structural damage. Detection metrics, however, reveal a safe operating region: within $\alpha \in [0, 0.10]$, F1 and mAP50 remain stable or slightly improve, while fusion quality steadily increases, achieving coordination between downstream tasks and fusion quality. Beyond $\alpha = 0.10$, detection accuracy drops sharply with mAP50 from 0.7238 to 0.6938, while fusion metrics continue to rise or plateau, revealing a clear divergence. This well-defined boundary corroborates prior findings Li et al. (2026a); Jiang et al. (2026); Chen et al. (2026b) on the inherent conflict between fusion quality and downstream task performance.

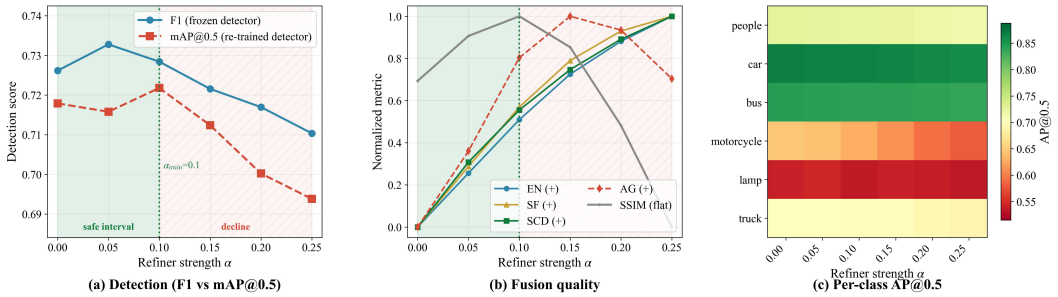

Figure 6: The Boundary-Aware Refiner reveals an inherent tension between fusion quality and downstream task performance, and identifies a safe operating region. (a) shows how downstream task metrics vary with refinement strength, (b) shows how fusion metrics vary with refinement strength, and (c) presents a heatmap of category-wise downstream task metrics across refinement strengths.We use a frozen detector to isolate the effect of α, and the full model in Table 3 corresponds to $\alpha = 0.10$

A.2 Implementation Details

A.2.1 Experiment Settings

Dataset Splits. We evaluate on four public benchmarks: M3FD (4,200 pairs) split into 2,940/420/840 for training, validation, and testing (7:1:2); FMB (1,500 pairs) into 1,070/150/280; MSRS (1,444 pairs) into 900/183/361; and RoadScene (221 pairs) into 100/50/71.

Main Framework Training. The main framework is optimized with AdamW, using an initial learning rate of 1×10^{-4}, decayed to 1×10^{-6} via cosine annealing, weight decay of 1×10^{-5}, and batch size of 8. We train for 100 epochs on all datasets. The total loss is a weighted combination of pixel loss, gradient loss, SSIM loss, VGG perceptual loss, degradation map anchor loss, edge gating loss, and spatial smoothness loss, with weights 1.0, 20.0, 2.0, 0.02, 0.5, 0.1, and 0.01, respectively. All experiments are conducted on a single NVIDIA RTX 5070 Ti (16 GB).

Refiner Training. The refiner is trained independently after the main framework converges, with all backbone parameters frozen. The optimizer is AdamW with an initial learning rate of 1×10^{-3}, decayed via cosine annealing to 1×10^{-5}. The refiner is trained for 150 additional epochs in all datasets. The residual scaling coefficient α is initialized to 0.1 and clipped to the range $[0, 0.1]$. The loss function includes pixel L1 loss, perceptual loss, edge loss, variance loss, and SCD loss, with weights 1.0, 0.1, 0.3, 0.1, and 0.1, respectively.

Ablation Studies and Boundary Exploration. Ablation experiments are conducted on a lightweight split of M3FD (200/50/70), with 50 training epochs and a fixed learning rate of 1×10^{-5}. We evaluate seven configurations: the full model, removal of the refiner, removal of the alignment encoder and infrared high-frequency injection, full removal of the gating map, and the gating map applied only to rank decomposition, only to QKV replacement, or only to the decoder. For the

independent-gate ablation, we additionally train two variants where the unified gating map is replaced by three independent $Conv1 \times 1$ adapters. Boundary exploration scans nine α values from 0 to 0.25 on 70 M3FD test images, with a finer step of 0.01 in the 0.08–0.12 interval, while the YOLO detector remains frozen without retraining.

Downstream Task Configuration. For object detection, we adopt YOLOv8n as the detector, trained on the M3FD fused images (3,360 pairs) following the official pipeline. For semantic segmentation, we use SegFormer-B3 as the segmenter, trained on the FMB (1,220 pairs) and MSRS (1,083 pairs) fused images with AdamW, learning rate of 1×10^{-4}, batch size of 4, and 80 epochs.

A.2.2 Training Strategy

Main framework loss. The training of DCFusion proceeds in two stages: main framework training and refiner training. The main framework is trained end-to-end with a multi-task objective covering pixel loss, gradient loss, SSIM loss, perceptual loss, and degradation anchor loss. Detailed definitions and weight configurations are provided in the supplementary material.

$$\mathcal{L}_{main} = \lambda_{pixel}\mathcal{L}_{pixel} + \lambda_{grad}\mathcal{L}_{grad} + \lambda_{ssim}\mathcal{L}_{ssim} + \lambda_{percep}\mathcal{L}_{percep} + \lambda_{deg}\mathcal{L}_{deg} \tag{21}$$

Pixel loss:

$$\mathcal{L}_{pixel} = \frac{1}{N}\|I_{fuse} - \max(I_{vi}, I_{ir})\|_1 \tag{22}$$

Gradient loss:

$$\mathcal{L}_{grad} = \frac{1}{N}\sum \|\max(|\nabla I_{vi}|, |\nabla I_{ir}|) - |\nabla I_{fuse}|\|_1 \tag{23}$$

SSIM loss:

$$\mathcal{L}_{ssim} = 1 - \mathrm{SSIM}(I_{fuse}, \max(I_{vi}, I_{ir})) \tag{24}$$

Perceptual loss:

$$\mathcal{L}_{percep} = \sum_{l\in\{3,4\}} \|\mathrm{VGG}_l(I_{fuse}) - \mathrm{VGG}_l(\max(I_{vi}, I_{ir}))\|_1 \tag{25}$$

Degradation anchor loss

$$\mathcal{L}_{deg} = \frac{1}{N}\|D_{vi} - D_{pseudo}\|_2^2 \tag{26}$$

Refiner loss. The refiner is trained independently after the main framework converges, with the main network parameters frozen. Its objective includes pixel loss, perceptual loss, and downstream task loss.

$$\mathcal{L}_{refiner} = \lambda_{pixel}^{ref}\|I_{refined} - I_{fuse}\|_1 + \lambda_{det}\mathcal{L}_{det} + \lambda_{percep}^{ref}\mathcal{L}_{percep}^{ref} \tag{27}$$

Where $\mathcal{L}_{det} = \mathcal{L}_{box} + \mathcal{L}_{cls} + \mathcal{L}_{obj}$ denotes the internal YOLO detection loss, and $\mathcal{L}_{percep}^{ref}$ shares the same definition as $\mathcal{L}_{percep}$ while operating on the refined image.

Notably, the downstream task model remains in inference mode throughout refiner training, providing only task-loss supervision without parameter updates. This design reflects our intention to explore the independent gain boundary of post-processing, rather than jointly optimizing with the downstream model, which lies beyond the scope of this work.

A.2.3 Knowledge Distillation Configuration

We distill the full model (1.46M backbone + 0.46M refiner) into DCFusion-Tiny (0.15M backbone + 0.1M refiner) with the teacher frozen. DCFusion-Tiny uses base width 24 and a shared siamese encoder (c/2→c→2c→c), retaining the degradation gate, frequency-rank decomposition, and dual cross-attention. TinyCNNRefiner uses $cond_{dim} = 16$, $feat = 32$, 5 blocks, and $\alpha \in [0, 0.1]$. Backbone distillation minimizes L1, SSIM, gradient, and gate MSE (weights 1, 2, 20, 5); refiner distillation uses L1, SSIM, and gradient. Both use 256×256 crops, AdamW (lr 1e-3, wd 1e-4), cosine annealing, and seed 42; backbone runs 100 epochs (batch 16), refiner 80 epochs (batch 8). The refiner acts as a quality equalizer, compensating the backbone compression loss.

Method	Params (M)	Speed (ms)
ReFusion	<u>0.06</u>	95.00
CoCoNet	9.17	53.10
EMMA	1.52	43.20
SAGE	0.14	**17.33**
DCEvo	2.00	1057.00
FSATFusion	**0.01**	59.40
UniFusion	5.44	499.09
CtrlFuse	6.30	694.00
HMMF	2.86	140.00
DAFusion	2.02	781.20
Ours	1.95	46.81
Ours (tiny)	0.25	<u>28.19</u>

Table 4: Model efficiency comparison.Best results are in **bold**, second best <u>underlined</u>.

A.2.4 Parameter Comparison

We compare the efficiency of our method with other approaches. Decision-Centric Fusion achieves a compelling balance between performance and efficiency. The full pipeline contains 1.95M parameters and runs at 46.81 ms per frame, delivering leading fusion quality and detection accuracy with far fewer parameters than DAFusion, which requires 2.02M parameters and 781.20 ms. With knowledge distillation, the tiny version reduces the pipeline to 0.25M parameters and 28.19 ms, achieving the fastest inference among all compared methods except SAGE while maintaining competitive performance. Although FSATFusion and SAGE are lighter, they fall substantially behind in fusion quality and downstream task performance. These results demonstrate that Decision-Centric Fusion delivers superior performance without sacrificing efficiency, and the distilled variant further extends its applicability to deployment scenarios with strict resource constraints.

A.3 Parameter Sensitivity Analysis

A.3.1 Experiment Settings

To evaluate the robustness of DCFusion to hyperparameter variations, we conduct a sensitivity analysis on three key parameters: the gradient loss weight λ_{grad},the degradation threshold τ, and the rank scaling factor. All experiments are performed on the lightweight M3FD split (200/50/70) for 50 epochs with a fixed learning rate of 1×10^{-5}.

	Fusion Metrics					**Detection Metrics**	
Variant	PSNR	EN	SF	AG	SCD	mAP50	mAP50-95
Default	<u>17.04</u>	6.80	<u>17.02</u>	6.63	1.45	0.708	0.482
$\lambda_{grad} = 10$	16.92	6.81	16.72	6.49	<u>1.46</u>	**0.714**	**0.486**
$\lambda_{grad} = 30$	16.83	<u>6.82</u>	16.93	6.61	1.45	0.710	<u>0.483</u>
$\tau = 0.3$	**17.12**	6.80	17.00	<u>6.64</u>	1.45	0.702	0.481
$\tau = 0.7$	16.75	**6.83**	17.00	<u>6.64</u>	**1.47**	<u>0.712</u>	<u>0.483</u>
Rank $0.5\times$	16.98	6.79	17.00	6.63	<u>1.46</u>	0.702	0.480
Rank $2\times$	17.01	6.78	**17.08**	**6.65**	<u>1.46</u>	0.707	0.481

Table 5: Parameter sensitivity analysis. Best results in **bold**, second best <u>underlined</u>.

A.3.2 Results and Analysis

The evaluation covers 8 fusion metrics and 5 detection metrics across seven configurations:$\lambda_{grad} \in \{10, 20, 30\}$,$\tau \in \{0.3, 0.5, 0.7\}$ and rank scaling $\in \{0.5\times, 1\times, 2\times\}$.

Among the tested parameters, λ_{grade} exhibits the most noticeable influence on performance. The configuration $\lambda_{grade} = 10$ achieves the highest mAP50, outperforming the default $\lambda_{grade} = 20$ by 0.8% . Higher gradient weight $\lambda = 30$ does not bring additional improvement, indicating that the

default setting slightly over-emphasizes edge sharpening. The degradation threshold $\tau = 0.7$, which adopts a more conservative strategy for infrared injection, yields mAP50=0.712, exceeding $\tau = 0.3$ by 1.4%, suggesting that cautious use of thermal information benefits detection performance. In contrast, varying the rank scaling factor from 0.5× to 2× leads to performance fluctuations below 1% across all metrics, confirming that the adaptive selection mechanism rather than the specific rank values is the key contributor.

Across all configurations, the maximum performance variation remains within 2%, demonstrating that DCFusion is highly robust to hyperparameter choices. Among the three factors, λ_{grade} is the only parameter with a notable impact, with its optimal value lower than the default, indicating that preserving visible texture smoothness is more beneficial than aggressive edge enhancement in degraded scenarios. The rank scaling factor shows negligible influence, validating the effectiveness of the adaptive rank selection mechanism. These results confirm that DCFusion maintains stable fusion quality and detection performance across a wide range of hyperparameter settings.

A.4 DOWANSTREAM TASKS PERFORMANCE

Table 6: Detection performance on M3FD dataset. Best results in **bold**, second best underlined.

Method	mAP50	mAP50-95	people	car	bus	motorcycle	lamp	truck
ReFusion	0.719	0.459	0.725	0.863	0.852	0.642	0.530	0.704
CoCoNet	0.610	0.401	0.689	0.831	0.786	0.276	0.454	0.626
EMMA	0.704	0.466	0.716	0.858	0.848	0.562	0.515	0.724
SAGE	0.717	0.476	**0.739**	0.860	0.861	0.588	0.525	0.730
DCEvo	0.715	0.473	0.426	0.605	0.722	0.259	0.301	0.526
FSATFusion	0.658	0.424	0.721	0.829	0.816	0.446	0.455	0.682
UniFusion	0.713	0.467	0.701	0.862	0.863	0.596	0.508	0.746
CtrlFuse	0.706	0.463	0.700	0.864	0.844	0.588	0.509	0.733
HMMF	0.699	0.463	0.592	0.866	0.838	0.633	0.462	0.800
DAFusion	0.721	**0.490**	0.720	0.868	0.847	0.578	0.527	**0.803**
Ours	**0.724**	0.478	0.718	**0.870**	0.848	**0.661**	**0.532**	0.715
Ours (tiny)	0.723	0.475	0.706	0.869	**0.865**	0.634	0.530	0.736

Table 7: Segmentation performance on FMB dataset. Best results in **bold**, second best underlined.

Method	mIoU	PA	mAcc	car	person	bike	guardrail	cone
ReFusion	0.351	0.832	0.485	0.742	0.247	0.669	0.738	0.891
CoCoNet	0.272	0.772	**0.968**	0.676	0.070	0.562	0.638	0.840
EMMA	0.341	0.819	0.411	0.721	0.243	0.631	0.718	0.876
SAGE	0.464	0.819	0.589	0.645	0.249	0.592	0.681	0.862
DCEvo	0.201	0.513	0.277	0.003	0.006	0.071	0.055	0.361
FSATFusion	0.462	0.816	0.583	0.644	0.263	0.593	0.678	0.860
UniFusion	0.478	0.825	0.604	0.663	0.267	0.604	0.678	0.867
CtrlFuse	0.436	0.793	0.572	0.583	0.047	0.501	0.586	0.861
HMMF	0.423	0.775	0.586	0.646	0.170	0.487	0.271	0.842
DAFusion	0.455	0.814	0.580	0.644	0.225	0.578	0.679	0.863
Ours	**0.507**	**0.847**	0.621	0.748	**0.416**	0.470	0.249	0.304
Ours (tiny)	0.503	0.841	0.621	**0.764**	0.284	**0.677**	**0.745**	**0.895**

Object Detection. As shown in the table, DCFusion achieves an mAP50 of 0.724 and an mAP50-95 of 0.478 on M3FD. At the category level, our method attains the highest precision on car (0.870), motorcycle (0.661), and street lamp (0.532), while also performing competitively on pedestrian and bus categories. Notably, DCFusion achieves the best results on both motorcycle and street lamp, the two categories with weak thermal radiation, confirming the enhancement effect of degradation-aware gating and infrared high-frequency injection on thermally faint targets. Detection on truck (0.715) is relatively lower, primarily because these objects frequently appear truncated

or at long distances in the dataset, leaving room for further improvement in coordinating visible details and infrared contours during fusion.

Semantic Segmentation. On FMB, DCFusion achieves the best performance across all three metrics, with mIoU of 0.507, PA of 0.847, and mAcc of 0.621. It excels on major categories such as car (0.748) and pedestrian (0.416), indicating that the joint texture-thermal enhancement strategy yields clear perceptual gains in densely semantic regions. However, on small or rare categories such as guardrail (0.249) and traffic cone (0.304), our segmentation accuracy falls behind some competing methods. This is primarily attributable to sample sparsity in the training data rather than a deficiency of the fusion method itself, as most competing methods also perform modestly on these two categories.

Overall, DCFusion performs best on texture-rich and thermally salient categories, yet its performance on rare or tiny categories is constrained by the input data, precisely reflecting the modality imbalance at the task level.

A.5 Discussion

A.5.1 Problem Formulation with Loss Decomposition

Recall the unified decision optimization problem in Eq. (1) of the main paper:

$$\min_{d_1,\ldots,d_n,g} \sum_{i=1}^{n} \mathcal{L}_i(d_i, g) + \lambda \Phi(g) \tag{28}$$

We decompose the loss terms $\mathcal{L}_i$ according to the three stages they regulate:

Encoding: $\mathcal{L}_{align}(d_1, g)$ is cross-modal alignment loss,d_1 controls the injection strength of infrared high-frequency features into the visible encoder.

Fusion: $\mathcal{L}_{rank}(d_2, g) + \mathcal{L}_{qkv}(d_3, g)$,$\mathcal{L}_{rank}$ supervises the adaptive rank selection based on degradation; $\mathcal{L}_{qkv}$ supervises the soft replacement of visible Query with infrared Query in cross-attention.

Decoding: $\mathcal{L}_{recon}(d_4, g)$ is reconstruction loss that enforces structural fidelity of the fused output.

The regularization term $\Phi(g)$ is defined as:

$$\Phi(g) = \|\nabla g\|_2^2 + \beta \cdot \mathcal{KL}(g \| g_0) \tag{29}$$

where $\|\nabla g\|_2^2$ is the spatial smoothness prior (encouraging adjacent pixels to share similar degradation decisions), and $\mathcal{KL}(g\|g_0)$ anchors the gate distribution to a neutral initialization $g_0 = 0.5$, preventing degenerate saturation. This design ensures that G remains both spatially consistent and numerically stable throughout training.

A.5.2 Analysis of the Unified Decision Variable

The decision variable G provides vertical scheduling but not semantic alignment. Resolving modality imbalance follows a strict causal chain: infrared high-frequency injection at the encoding stage is the prerequisite, without which later modules lack texture to process; adaptive high-low separation is the means, as otherwise weak signals are masked by noise; and soft QKV replacement is the safeguard, as the visible query remains noisy. These three modules are interlocking, each mathematically compensating for the deficiencies of its predecessor.Here is the detailed analysis.

The unified gate G is embedded into the gradient paths of three submodules: rank selection $\mathcal{R}$, QKV replacement $\mathcal{Q}$, and fusion weighting $\mathcal{W}$. The total gradient w.r.t. G is:

$$\frac{\partial \mathcal{L}_{total}}{\partial G} = \frac{\partial \mathcal{L}_{rank}}{\partial G} + \frac{\partial \mathcal{L}_{qkv}}{\partial G} + \frac{\partial \mathcal{L}_{fuse}}{\partial G} \tag{30}$$

For stage-wise independent gates G_1, G_2, G_3, the gradient for each is:

$$\frac{\partial \mathcal{L}_{total}}{\partial G_i} = \frac{\partial \mathcal{L}_i}{\partial G_i}, \quad i = 1, 2, 3 \tag{31}$$



Figure 7: Performance of decision gate on actual degradation conditions.Performance of decision gate on actual degradation conditions. Quantitative validation is provided in Appendix A.6.

The unified design differs in that G's update reflects a weighted consensus of gradient signals from all three modules, each module pulls G in a direction that benefits its own objective, and the shared variable finds a joint optimum that no independent gate can reach. This is empirically validated in Table 3 of the main paper: replacing unified G with three independent gates drops mAP from 0.724 to 0.704.

The gradient paths to each module are as follows:

To $\mathcal{R}$: G controls the mixing weight between low-rank and high-rank branches (Eq. 10-11, main paper); $\partial\mathcal{L}_{rank}/\partial G$ flows through the rank mixing coefficients to favor high-rank branches in degraded regions.

To $\mathcal{Q}$: G determines the soft replacement weight θ (Eq. 15, main paper); $\partial\mathcal{L}_{qkv}/\partial G$ flows through θ to adjust the infrared Query ratio.

To $\mathcal{W}$: G is downsampled to produce the fusion weight map W_{fuse} (Eq. 18, main paper); $\partial\mathcal{L}_{fuse}/\partial G$ flows through the weight generation network to adjust modality contributions per spatial location.

Thus, a single update to G simultaneously receives supervisory signals from rank selection, attention, and fusion weighting, ensuring that no module optimizes in isolation and preventing the cross-stage inconsistency that stage-wise designs inherently suffer from.

To further verify that the unified decision gate G does not impose blind over-coupling but instead adapts to actual degradation conditions, we visualize G on representative samples from M3FD under varying degradation levels (Figure 7). In scenes with fog or low illumination, G assigns high activation values to regions where visible information is severely degraded, such as distant objects, sky areas, and shadowed regions, while keeping low activation in well-lit foreground areas. In clear daytime scenes, G remains largely uniform with moderate values, indicating that the decision variable does not force unnecessary infrared dominance when visible textures are reliable. These observations confirm that G functions as a degradation-aware decision hub rather than a rigid coupling mechanism: its spatial distribution directly reflects the degradation conditions, ensuring that

modality decisions are made on a per-region basis. This adaptive behavior mitigates the risk of over-coupling and supports the claim that the unified gate enables spatially differentiated fusion decisions across the full pipeline.

A.5.3 Self-Adaptive Rank Selection

High-frequency components, carrying edges, textures, and small-target contours, inherently require higher rank and are set with $r_l = 32$ and $r_h = 64$. Low-frequency components, dominated by backgrounds and large-scale structures, have lower intrinsic rank and use $r_l = 4$ and $r_h = 16$. This design follows the rate-distortion principle: high-entropy details demand more channels to maintain fidelity, while low-entropy structures tolerate fewer channels without significant information loss. The rate-distortion curve for each frequency band is estimated empirically from the training set—the rank thresholds are chosen at the knee point where increasing rank yields diminishing returns in feature reconstruction quality. This analysis confirms that the specific rank values are not arbitrary hyperparameters but are grounded in the information-theoretic properties of the data distribution, as validated by the parameter sensitivity experiments, where scaling ranks by $0.5\times$ or $2\times$ results in less than 1% performance variation, confirming that the adaptive selection mechanism is the key contributor rather than the absolute rank values.

A.6 Degradation-Aware Gating Validation

The degradation estimator takes only the visible image as input:$D_{vi} = \sigma(\mathrm{CNN}_{deg}(I_{vi}))$. Infrared is used only during training to construct a self-supervised pseudo label:

$$D_{pseudo} = D_{grad} \odot \max(S_{ir}^{norm}, \beta) \tag{32}$$

where D_{grad} measures visible gradient sparsity and S_{ir}^{norm} is normalized infrared local contrast saliency. The estimator is supervised by

$$L_{deg} = \frac{1}{N} \|D_{vi} - D_{pseudo}\|_2^2 \tag{33}$$

No external model, dataset, or annotation is used. At inference, both D_{vi} and G depend only on the visible image.

To verify that the shared decision variable g reflects visible degradation rather than trivial brightness or haze, we conduct two analyses: (i) correlation with degradation proxies, (ii) stratified metrics under different degradation levels.

Table 8: Correlation between mean(G) and degradation proxies.

Proxy	Pearson	Spearman
Gradient	-0.985	-0.970
Local contrast	-0.942	-0.946
D_{pseudo}	$+0.985$	$+0.955$
Brightness	-0.301	-0.151
Dark channel	$+0.081$	$+0.030$

Correlation with degradation proxies. We compute Pearson and Spearman correlations between $\mathrm{mean}(G)$ and several proxies on the M3FD test set. As shown in Table 8, G is strongly negatively correlated with gradient magnitude (Pearson (=-0.985), Spearman (=-0.970)) and local contrast (Pearson (=-0.942), Spearman (=-0.946)), and strongly positively correlated with the training anchor D_{pseudo} (Pearson (=+0.985), Spearman (=+0.955)). In contrast, correlations with brightness (Pearson (=-0.301), Spearman (=-0.151)) and dark channel (Pearson (=+0.081), Spearman (=+0.030)) are near zero.

These results indicate that G captures texture sparsity rather than simple illumination or haze. The near-zero correlation with brightness and dark channel is particularly important: it shows that G is not a trivial brightness or transmission estimator.

Stratified metrics. We stratify M3FD test images into four equal-size groups according to D_{pseudo} quartiles, corresponding to clear, mildly degraded, moderately degraded, and heavily degraded

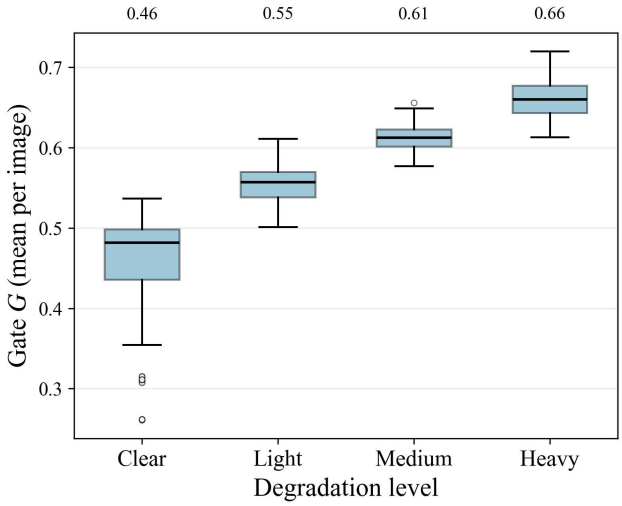

Figure 8: Distribution of the shared decision signal G across degradation levels.Test images are divided into four equal-size groups by D_{pseudo} quartiles. The boxplots show that G increases monotonically from clear to heavy degradation, with clearly separated interquartile ranges. This confirms that G reflects degradation severity and that the pseudo-label stratification is meaningful.

scenes. Figure 9 shows the distribution of the shared decision signal G across these four levels. The median of G increases monotonically from clear to heavily degraded scenes, and the interquartile ranges of adjacent levels are largely separated. This trend confirms that D_{pseudo} provides a meaningful stratification of degradation severity, and that G consistently responds to this stratification by assigning higher values to more degraded regions. Together with the correlation analysis in Table 8, these results support that G functions as a degradation-aware texture-sparsity estimator, rather than a trivial brightness or haze map.

Intuitively, as visible texture becomes sparser under adverse conditions, D_{pseudo} increases; G inherits this trend and therefore provides a spatially varying condition that indicates where modality-dependent processing should be strengthened.

Overall, these analyses show that G is a degradation-aware texture-sparsity estimator, not a trivial brightness or haze map, and that sharing it across the main fusion submodules provides measurable benefits.

A.7 Main Results

For better readability, we provide an enlarged version of the main comparison results in this appendix. The main paper reports all results in a single compact table due to the page limit. Here, we split the same data into two tables with larger fonts and wider spacing. The values are identical to those in Table 1 of the main paper.

A.8 Robustness to Random Seeds

To address the concern about single-run evaluation, we repeat the main experiments with three random seeds (42, 100, 2025) and report the mean and standard deviation in Tables 11. Across both M3FD and FMB, the standard deviations are consistently small (below 1% of the mean for all metrics), indicating that the reported performance is stable and not sensitive to initialization. This further supports that the gains of DCFusion are architecture-driven rather than the result of a lucky seed.

Table 9: Comparison on M3FD and FMB datasets (PSNR, EN, SF, AG, SCD). Best results in **bold**, second best underlined.

	M3FD					FMB				
Method	PSNR	EN	SF	AG	SCD	PSNR	EN	SF	AG	SCD
ReFusion	16.13	6.91	16.00	6.29	1.60	17.36	6.81	15.17	5.42	1.52
CoCoNet	15.94	6.67	8.32	3.21	1.50	17.30	6.50	7.67	2.61	1.39
EMMA	16.21	6.92	15.87	6.51	1.51	17.77	6.69	14.99	5.54	1.51
SAGE	14.82	6.89	13.37	5.26	**1.71**	15.75	**6.84**	12.64	4.53	1.53
DCEvo	15.90	6.90	14.84	5.71	1.52	17.19	6.69	13.76	4.83	1.50
FSATFusion	15.83	6.70	11.93	4.50	1.31	17.28	6.52	11.88	4.17	1.29
UniFusion	16.31	6.88	14.75	5.81	1.46	18.01	6.65	14.06	4.97	1.45
CtrlFuse	16.28	6.83	14.90	6.05	1.47	18.00	6.62	13.75	5.08	1.39
HMMF	16.17	6.64	13.66	5.47	0.91	17.98	6.41	13.46	4.80	0.65
DAFusion	15.67	6.74	10.41	4.22	**1.71**	16.97	6.68	9.92	3.60	1.56
Ours	16.35	**6.98**	**19.49**	**7.90**	1.62	**18.03**	6.70	**15.39**	**5.58**	**1.59**
Ours (tiny)	**16.64**	6.92	19.09	7.71	1.62	17.99	6.65	15.11	5.44	1.57

Table 10: Comparison on MSRS and RoadScene datasets (PSNR, EN, SF, AG, SCD). Best results in **bold**, second best underlined.

	MSRS					RoadScene				
Method	PSNR	EN	SF	AG	SCD	PSNR	EN	SF	AG	SCD
ReFusion	20.39	6.62	**11.63**	**4.39**	1.67	15.67	7.18	6.98	2.97	**1.67**
CoCoNet	17.80	5.37	7.34	2.53	1.24	15.46	6.94	10.72	4.84	1.30
EMMA	20.44	6.74	11.55	4.38	1.63	14.69	7.22	7.13	3.20	1.65
SAGE	19.57	6.02	10.42	3.70	1.42	14.03	7.06	10.53	4.72	1.50
DCEvo	20.15	6.64	11.46	4.32	1.67	15.72	7.21	13.01	5.84	1.46
FSATFusion	20.11	6.60	11.38	4.19	1.54	15.56	6.92	5.69	2.16	1.21
UniFusion	16.38	6.68	11.16	4.14	1.71	15.65	6.98	4.76	2.24	1.39
CtrlFuse	20.24	6.43	10.49	4.11	1.49	15.38	7.12	10.66	4.90	1.37
HMMF	12.91	6.06	7.91	2.84	0.89	12.35	6.91	11.27	4.75	0.59
DAFusion	18.74	5.81	8.24	3.03	1.31	15.52	7.12	7.17	3.26	1.62
Ours	**20.88**	**6.76**	11.09	4.18	**1.79**	**16.20**	**7.28**	**14.52**	**6.92**	1.33
Ours (tiny)	20.58	6.69	10.74	3.98	1.68	16.18	7.27	13.18	6.20	1.35

Table 11: Mean $\pm$ std over three random seeds (42, 100, 2025) on M3FD and FMB.

Dataset	Metric	Mean $\pm$ Std
M3FD	PSNR	16.453 ± 0.080
	EN	6.968 ± 0.007
	SF	19.460 ± 0.020
	AG	7.897 ± 0.007
	SCD	1.611 ± 0.015
	mAP50	0.7239 ± 0.0007
FMB	PSNR	17.703 ± 0.096
	EN	6.712 ± 0.003
	SF	15.426 ± 0.013
	AG	5.583 ± 0.002
	SCD	1.562 ± 0.004
	mIoU	0.4707 ± 0.0010