PDD-FUSION: PHYSICS-GUIDED DEGRADATION-DECOUPLED LEARNING FOR LOW-LIGHT INFRARED-VISIBLE IMAGE FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-light infrared–visible image fusion is challenged by heterogeneous degradation effects across modalities, spatial regions, and frequency components. Visible low-frequency information can suffer from poor exposure, high-frequency responses may mix useful structures with noise, and infrared observations may exhibit weak local thermal variations. We propose PDD-Fusion, a Physics-guided Degradation-Decoupled framework that explicitly models degradation-related reliability and assigns different degradation factors to specialized fusion operations. Specifically, seven imaging-inspired priors are organized into exposure, high-frequency reliability, and infrared thermal groups, and encoded into spatial and global degradation prompts. Guided by these representations, PDD-Fusion performs high-frequency reliability modulation, low-frequency exposure-aware fusion, and gated infrared residual compensation while retaining direct modality paths for reconstruction. Trained only on LLVIP, PDD-Fusion achieves leading performance on multiple fusion metrics and generalizes effectively to M^3FD and MSRS without fine-tuning, demonstrating the effectiveness of degradation-specific fusion under low-light conditions.

1 INTRODUCTION

Infrared–visible image fusion (IVIF) aims to integrate complementary information from heterogeneous sensors into a unified representation. Infrared images can preserve thermal responses under poor illumination, whereas visible images provide richer texture and structural details. Recent deep fusion methods have improved feature representation and cross-modal interaction, ranging from CNN-based approaches (Li & Wu, 2018; Xu et al., 2020) to transformer-based and task-oriented frameworks (Ma et al., 2022; Tang et al., 2022a; Zhao et al., 2023a). Despite these advances, most fusion methods focus on improving information aggregation, while the reliability of source observations is often modeled only implicitly. This becomes problematic in low-light scenes, where the usefulness of multimodal information varies across spatial regions and frequency components.

Low-light degradation is heterogeneous rather than uniform. Insufficient exposure mainly affects slowly varying visible intensity information, whereas high-frequency responses may contain both useful structures and noise amplified by poor illumination. Infrared observations are more robust to illumination changes, but locally flat thermal regions may provide weak complementary information. Consequently, useful scene content and degradation-sensitive responses can become entangled during feature extraction and multimodal interaction. As illustrated in Fig. 1, conventional fusion pipelines typically extract modality features and directly aggregate them, which may propagate unreliable responses and result in noise leakage, detail loss, or exposure distortion. This suggests that low-light IVIF is not only an information aggregation problem, but also an information reliability selection problem: different degradation factors should be handled according to their modality dependencies and frequency characteristics.

Degradation-aware modeling has been explored in both image restoration and IVIF. Restoration methods demonstrate that degradation representations can effectively condition adaptive processing (Chen Wei, 2018; Li et al., 2022; Potlapalli et al., 2023), while fusion methods incorporate low-light enhancement, semantic guidance, or modality-specific priors (Tang et al., 2023; Chen et al., 2024;

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.

Yi et al., 2024; Li et al., 2026; Jiang et al., 2026; Zhang et al., 2026). These studies demonstrate the importance of degradation information and prior knowledge for robust multimodal fusion. However, degradation cues are often used as auxiliary guidance for generic enhancement or feature interaction, while the correspondence between heterogeneous degradation characteristics and appropriate fusion operations is less explicitly modeled. In particular, degradation factors with different frequency characteristics may still be handled within a shared interaction mechanism.

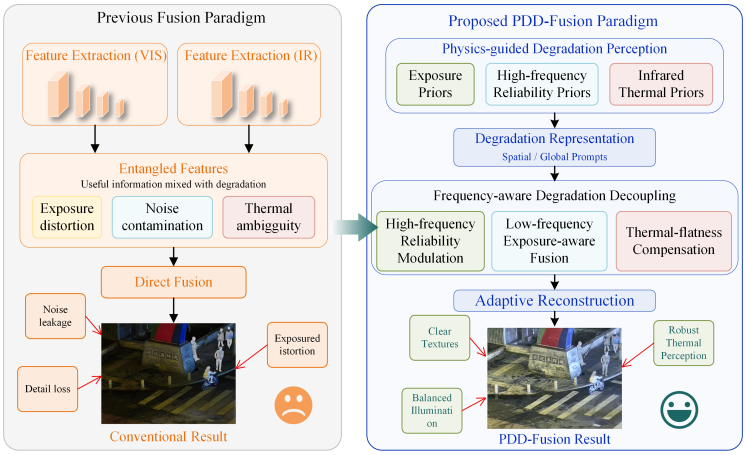

Figure 1: Motivation of PDD-Fusion. Existing methods suffer from degradation entanglement during implicit feature learning, while PDD-Fusion introduces degradation-aware perception and frequency-specific processing.

To address this issue, we propose PDD-Fusion, a Physics-guided Degradation-Decoupled framework for low-light infrared–visible image fusion. As shown in the right part of Fig. 1, PDD-Fusion explicitly introduces degradation perception before multimodal fusion. Specifically, seven source-derived priors are constructed to characterize visible exposure conditions, high-frequency reliability, and infrared thermal characteristics, and are encoded into spatial and global degradation prompts. Based on their modality dependencies and frequency characteristics, different degradation factors are assigned to three specialized processing pathways: visible high-frequency reliability modulation, low-frequency exposure-aware modality mixing, and gated infrared residual compensation. We refer to this degradation-specific operational separation as degradation decoupling, which prevents heterogeneous degradation effects from being handled solely through a shared fusion operator.

The resulting degradation-aware features are further modulated by global imaging conditions and integrated with direct modality features for adaptive reconstruction. In this way, PDD-Fusion establishes an explicit degradation-perception–representation–decoupling pipeline while retaining complementary source information. We evaluate PDD-Fusion on LLVIP (Jia et al., 2021), M^3FD (Liu et al., 2022), and MSRS (Tang et al., 2022b). Trained only on LLVIP, the model achieves leading performance on multiple fusion metrics and transfers effectively to M^3FD and MSRS without fine-tuning, demonstrating the effectiveness and cross-dataset generalization of the proposed degradation-specific fusion strategy.

The main contributions of this work are summarized as follows:

- We formulate low-light infrared–visible fusion from a degradation-specific reliability perspective, explicitly associating heterogeneous degradation factors with their modality dependencies and frequency characteristics.
- We introduce a physics-guided degradation representation composed of seven source-derived priors and encode them into complementary spatial and global prompts for adaptive degradation-aware feature processing.
- We develop a degradation-decoupled fusion architecture that assigns different degradation factors to dedicated frequency-aware processing pathways while preserving direct modality information for adaptive reconstruction.

2 Related Work

2.1 Infrared–Visible Image Fusion

Infrared–visible image fusion (IVIF) integrates complementary thermal and appearance information from heterogeneous sensors. Early deep learning methods employed encoder–decoder architectures to extract and combine modality features (Li & Wu, 2018; Li et al., 2021). Subsequent studies developed unified learning objectives and transformer-based interactions to improve information preservation and cross-modal modeling (Xu et al., 2020; Ma et al., 2022). CDDFuse introduced correlation-driven feature decomposition, using separate paths for low-frequency global information and high-frequency local information while modeling shared and modality-specific representations (Zhao et al., 2023a). Other methods incorporate target cues or semantic supervision so that fusion can further benefit downstream perception (Liu et al., 2022; Tang et al., 2022a).

These methods establish effective mechanisms for complementary feature representation and multimodal interaction. However, their fusion strategies mainly focus on how information should be combined, while the reliability of the source information itself is often handled implicitly. Under low-light conditions, visible exposure, high-frequency reliability, and infrared thermal contrast exhibit different modality dependencies and frequency characteristics. This motivates fusion mechanisms that explicitly account for heterogeneous information reliability before multimodal aggregation.

2.2 Degradation-Aware and Prior-Guided Fusion

Image restoration has extensively explored degradation-aware processing. Retinex-based and zero-reference methods address illumination and exposure degradation in single images (Chen Wei, 2018; Guo et al., 2020), while AirNet and PromptIR learn degradation-related representations to condition adaptive restoration (Li et al., 2022; Potlapalli et al., 2023). In IVIF, methods designed for low-light scenes combine visible enhancement with multimodal fusion (Tang et al., 2023; Chen et al., 2024; Xie et al., 2025). For example, DCDFusion couples adaptive visible enhancement with dual-consistency feature learning and adaptive fusion (Xie et al., 2025).

Other approaches introduce additional guidance beyond conventional fusion features. Text-IF employs textual information for degradation-aware fusion (Yi et al., 2024); MdaIF uses language-driven semantics to address multiple degradations (Li et al., 2026); and MMAIF combines language guidance with an all-in-one framework spanning multiple fusion tasks and degradation settings (Cao et al., 2025). Generative methods instead incorporate restoration priors: DRMF exploits modality-specific diffusion priors (Tang et al., 2024), while ReCoFuse couples modality restoration and fusion throughout the diffusion process (Zhang et al., 2026). PhyFusion derives modality-specific components from radiative-transfer and Retinex models and imposes physical consistency for prior-guided fusion (Jiang et al., 2026).

These studies demonstrate that degradation information, semantic guidance, and physical or generative priors can effectively improve multimodal fusion. However, such guidance is commonly used for generic enhancement, feature modulation, or restoration-aware interaction. The correspondence between heterogeneous degradation factors and the specific fusion operations required to handle them remains less explicitly modeled. PDD-Fusion addresses this issue from a degradation-specific reliability perspective: source-derived priors characterize exposure, high-frequency reliability, and infrared thermal conditions, which are subsequently assigned to dedicated frequency-aware processing pathways before multimodal reconstruction.

3 Method

PDD-Fusion performs degradation-specific processing before multimodal reconstruction. Given paired infrared and visible intensity images $I_{ir}, I_{vi} \in [0,1]^{H\times W}$, PDD-Fusion predicts a fused intensity image

$$I_f = f_\theta(I_{ir}, I_{vi}) \in [0,1]^{H\times W}. \tag{1}$$

As shown in Fig. 2, the framework consists of three stages: physics-guided degradation perception, frequency-aware degradation decoupling, and adaptive reconstruction. First, seven source-derived priors describing visible exposure, high-frequency reliability, and infrared thermal characteristics

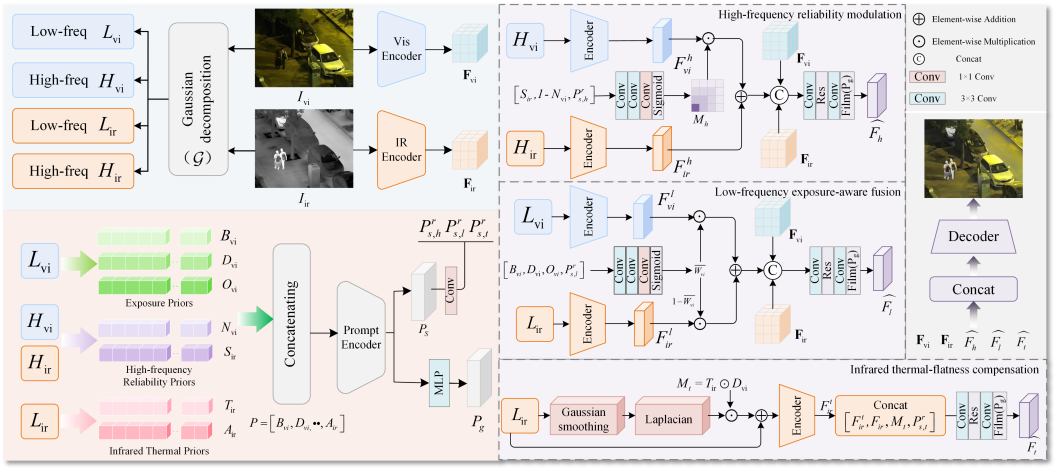

Figure 2: Overview of PDD-Fusion. Source-derived priors are encoded into spatial and global degradation prompts, which guide three frequency-aware processing branches. Their outputs are integrated with direct modality features for adaptive reconstruction.

are encoded into spatial and global degradation prompts. The spatial prompt guides three specialized branches for high-frequency reliability modulation, low-frequency exposure-aware fusion, and gated infrared residual compensation, while the global prompt modulates their outputs according to the overall imaging condition. Finally, the degradation-aware branch features are combined with direct modality features to reconstruct I_f, preserving complementary source information during degradation-specific processing. The network is trained without reference fused images using the reliability-aware objectives in Sec. 3.4.

3.1 Physics-Guided Degradation Perception

To characterize heterogeneous source reliability before multimodal interaction, we construct a seven-channel degradation representation organized into three groups: visible exposure reliability, high-frequency reliability, and infrared thermal reliability. These source-derived priors are jointly encoded into spatial and global degradation prompts for subsequent frequency-aware processing.

Frequency decomposition and shared operators. For each modality $m \in \{\mathrm{ir}, \mathrm{vi}\}$, we decompose the input into low- and high-frequency components:

$$L_m = \mathcal{G}_{7,1.5}(I_m), \qquad H_m = I_m - L_m, \tag{2}$$

where $\mathcal{G}_{k,\sigma}$ denotes Gaussian filtering with a $k \times k$ kernel and standard deviation σ. For a map X, we normalize each sample and channel spatially by

$$d(X) = \begin{cases} r(X), & r(X) \geq \epsilon, \\ 1, & r(X) < \epsilon, \end{cases} \qquad \mathcal{N}(X) = \mathrm{clip}_{[0,1]}\left(\frac{X - \min X}{d(X)}\right), \tag{3}$$

where $r(X) = \max X - \min X$ and $\epsilon = 10^{-6}$. Local variation is measured by

$$\mathrm{Std}_k(X) = \sqrt{\max\left(\mathbb{E}_k[X^2] - \mathbb{E}_k[X]^2,\, 0\right) + \epsilon}, \tag{4}$$

where $\mathbb{E}_k$ denotes local averaging with a $k \times k$ box window. We write $\mathrm{sigm}(x) = 1/(1+\exp(-x))$.

Visible exposure-reliability priors. Since illumination degradation mainly affects slowly varying visible intensity, we derive relative exposure cues from L_{vi}:

$$B_{vi} = \mathcal{N}(L_{vi}), \qquad D_{vi} = 1 - B_{vi}, \qquad O_{vi} = \mathrm{sigm}(20(B_{vi} - 0.85)). \tag{5}$$

B_{vi} and D_{vi} describe relatively bright and dark regions, respectively, while O_{vi} emphasizes high relative brightness. Together, they provide spatial exposure conditions for low-frequency modality selection.

High-frequency reliability priors. Visible high-frequency responses may contain both useful structures and noise-sensitive fluctuations. We therefore construct a darkness-weighted variation cue:

$$R_n = \mathcal{N}(\mathrm{Std}_7(H_{vi})), \qquad N_{vi} = \mathcal{N}(\mathcal{G}_{3,0.8}\left[R_n \odot (0.5 + 0.5D_{vi})\right]). \tag{6}$$

The darkness term increases sensitivity to unstable variation in poorly illuminated regions. Since valid texture can also produce local variation, N_{vi} is treated as a reliability cue rather than an explicit noise label.

To provide complementary infrared structural evidence, we compute a smoothed structure tensor from H_{ir}:

$$\mathbf{J}_{ir} = \mathcal{G}_{5,1.0}\left(\begin{bmatrix} H_x^2 & H_xH_y \\ H_xH_y & H_y^2 \end{bmatrix}\right), \tag{7}$$

where H_x and H_y are Sobel responses. Let $\lambda_1 \geq \lambda_2$ be the eigenvalues of $\mathbf{J}_{ir}$. We define

$$C_{ir} = \frac{\lambda_1 - \lambda_2}{\lambda_1 + \lambda_2 + \epsilon}, \qquad S_{ir} = \mathcal{N}(\mathcal{G}_{3,0.8}\left[C_{ir} \odot \mathcal{N}(\lambda_1 + \lambda_2)\right]), \tag{8}$$

where coherence emphasizes directional structure and the energy term suppresses weak high-frequency responses. Thus, N_{vi} and S_{ir} provide complementary high-frequency reliability evidence.

Infrared thermal-reliability priors. We characterize infrared low-frequency flatness by

$$T_{ir} = \mathrm{clip}_{[0,1]}\left\{\mathcal{G}_{3,0.8}\left[1 - \mathcal{N}(\mathrm{Std}_9(L_{ir}))\right]\right\}, \tag{9}$$

where larger T_{ir} indicates weaker local variation. We further extract positive local infrared contrast using a center–surround response:

$$\bar{I}_{ir} = \mathcal{N}(I_{ir}), \qquad A_{ir} = \mathrm{sigm}\big(10\left\{\mathcal{N}\big(\mathrm{ReLU}[\bar{I}_{ir} - \mathcal{G}_{15,3.0}(\bar{I}_{ir})]\big) - 0.3\right\}\big). \tag{10}$$

T_{ir} and A_{ir} respectively describe local flatness and positive contrast, providing complementary cues for infrared reliability.

Finally, the seven priors are concatenated as

$$\mathbf{P} = [B_{vi}, D_{vi}, O_{vi}, N_{vi}, S_{ir}, T_{ir}, A_{ir}], \tag{11}$$

and encoded into spatial and global degradation prompts for the subsequent frequency-aware decoupling stage.

3.2 Frequency-Aware Degradation Decoupling

The degradation representation $\mathbf{P}$ conditions three specialized processing pathways according to their modality dependencies and frequency characteristics. A convolutional encoder first produces a spatial prompt, from which an image-level prompt is derived:

$$\mathbf{P}_s = \mathcal{E}_s(\mathbf{P}), \qquad \mathbf{p}_g = \mathrm{MLP}(\mathrm{GAP}(\mathbf{P}_s)), \tag{12}$$

where $\mathbf{P}_s$ has 64 channels and $\mathbf{p}_g$ has dimension 64. For each branch $b \in \{h, l, t\}$, a learned projection $\mathbf{P}_{s,b}^r = \mathcal{Q}_b(\mathbf{P}_s)$ reduces the spatial prompt to 16 channels. Shallow modality-specific encoders further extract $F_{ir}, F_{vi} \in \mathbb{R}^{B\times C\times H\times W}$ with $C = 48$.

High-frequency reliability modulation. Visible high-frequency responses are regulated using infrared structure, visible reliability, and the branch-specific spatial prompt:

$$M_h = \mathrm{sigm}\big(\phi_h\big[S_{ir},\, 1 - N_{vi},\, \mathbf{P}_{s,h}^r\big]\big), \tag{13}$$

where ϕ_h predicts a spatial reliability gate. The encoded visible high-frequency feature is then modulated as

$$\widetilde{F}_{vi}^h = M_h \odot \mathcal{E}_{vi}^h(H_{vi}). \tag{14}$$

A fusion block combines F_{ir}, F_{vi}, $\mathcal{E}_{ir}^h(H_{ir})$, and $\widetilde{F}_{vi}^h$ to obtain F_h. This preserves infrared and direct modality information while spatially regulating unreliable visible high-frequency responses.

Low-frequency exposure-aware fusion. For low-frequency information, we predict a spatially adaptive visible contribution from both modalities, the exposure priors, and the projected spatial prompt:

$$\mathbf{Z}_l = [L_{ir}, L_{vi}, B_{vi}, D_{vi}, O_{vi}, \mathbf{P}^r_{s,l}], \qquad \overline{W}_{vi} = \mathrm{sigm}(\phi_l(\mathbf{Z}_l)) \,. \tag{15}$$

The visible weight is further regulated by the high-relative-brightness cue:

$$W_{vi} = \mathrm{clip}_{[0,1]} \left[\overline{W}_{vi} \odot (1 - O_{vi})\right], \qquad W_{ir} = 1 - W_{vi}, \tag{16}$$

and the mixed low-frequency feature is

$$F^l_{\mathrm{mix}} = W_{vi} \odot \mathcal{E}^l_{vi}(L_{vi}) + W_{ir} \odot \mathcal{E}^l_{ir}(L_{ir}). \tag{17}$$

The fusion block combines F^l_{mix} with the unweighted low-frequency features and F_{vi}, F_{ir} to produce F_l, retaining source information alongside exposure-aware modality selection.

Gated infrared residual compensation. To identify regions where visible information is weak and the corresponding infrared response is locally flat, we define a compensation-need map

$$M_t = D_{vi} \odot T_{ir}. \tag{18}$$

This mask controls a local second-order adjustment to the infrared low-frequency component:

$$\widetilde{L}_{ir} = \mathrm{clip}_{[0,1]} \left[L_{ir} + \alpha M_t \odot \Delta(\mathcal{G}_{5,1.0}(L_{ir}))\right], \qquad \alpha = 0.5, \tag{19}$$

where Δ denotes the reflection-padded four-neighbor discrete Laplacian. Since a constant neighborhood produces zero Laplacian response, the operation enhances only existing local variations rather than creating missing thermal information. The compensated component $\widetilde{L}_{ir}$ is encoded and combined with F_{ir}, M_t, and $\mathbf{P}^r_{s,t}$ to obtain F_t.

Finally, branch-specific projections of the global prompt generate channel-wise modulation parameters:

$$\widehat{F}_b = (1 + \boldsymbol{\gamma}_b) \odot F_b + \boldsymbol{\beta}_b, \qquad b \in \{h, l, t\}. \tag{20}$$

The spatial prompt therefore controls local degradation-specific processing, while the global prompt adapts the three branch outputs to the overall imaging condition, realizing frequency-aware degradation decoupling before reconstruction.

3.3 Adaptive Reconstruction and Output Representation

After frequency-aware degradation decoupling, the three branch features are combined with direct infrared and visible features:

$$F_{\mathrm{cat}} = \left[F_{ir},\, F_{vi},\, \widehat{F}_h,\, \widehat{F}_l,\, \widehat{F}_t\right] \in \mathbb{R}^{B \times 5C \times H \times W}, \tag{21}$$

where $C = 48$. Retaining F_{ir} and F_{vi} preserves complementary source information that may be attenuated during degradation-specific processing.

The decoder first projects the concatenated $5C$ channels to C channels using a 3×3 convolution, followed by four residual blocks, a bottleneck reducing the width to $C/2$, and a final 3×3 convolution. The fused intensity is predicted as

$$I_f = \mathrm{sigm}\left(\mathcal{D}_\theta(F_{\mathrm{cat}})\right), \tag{22}$$

where the sigmoid constrains the output to $[0, 1]$.

For RGB visible inputs, I_{vi} denotes the luminance channel Y_{vi}. PDD-Fusion predicts only the fused luminance I_f, which is combined with the visible chrominance channels Cb_{vi} and Cr_{vi} through inverse YCbCr-to-RGB conversion for color visualization.

3.4 Reliability-Aware Training Objectives

Without reference fused images, PDD-Fusion is optimized using source-derived objectives for intensity, structure, and reliable high-frequency preservation.



Figure 3: Qualitative comparisons between Ours and state-of-the-art methods on LLVIP, M^{3}FD, and MSRS datasets.

Exposure-adaptive intensity preservation. We adapt the visible intensity according to the relative-brightness prior:

$$\gamma = 0.4 + 0.6B_{vi}, \qquad I_{\text{tgt}} = \max(I_{ir}, I_{vi}^{\gamma}), \tag{23}$$

where all operations are element-wise. Smaller B_{vi} produces stronger adjustment in relatively dark visible regions, while the maximum operation retains stronger infrared responses. The intensity loss is

$$\mathcal{L}_{\text{int}} = \frac{1}{HW} \left\| I_f - I_{\text{tgt}} \right\|_1 . \tag{24}$$

Structural consistency. We further preserve local structure using

$$\mathcal{L}_{\text{ssim}} = 1 - \text{SSIM}_7(I_f, I_{\text{tgt}}), \tag{25}$$

where SSIM_7 uses a 7×7 local window.

Reliability-aware gradient preservation. To reduce noise-sensitive visible gradients, the visible contribution is weighted by $1 - N_{vi}$. The gradient magnitude is

$$g(X) = \sqrt{(K_x * X)^2 + (K_y * X)^2 + \epsilon}, \tag{26}$$

where K_x and K_y are Sobel kernels and $\epsilon = 10^{-6}$. We define

$$\mathcal{L}_{\text{grad}} = \frac{1}{HW} \left\| g(I_f) - \max\left(g(I_{ir}), (1 - N_{vi}) \odot g(I_{vi})\right) \right\|_1 . \tag{27}$$

Here, N_{vi} acts as a soft reliability prior rather than an explicit noise label.

The total objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{int}} + 2\mathcal{L}_{\text{ssim}} + 5\mathcal{L}_{\text{grad}}, \tag{28}$$

with all losses averaged over the training batch.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate PDD-Fusion on LLVIP (Jia et al., 2021), MSRS (Tang et al., 2022b), and M^{3}FD (Liu et al., 2022). LLVIP contains 15,488 aligned infrared–visible image pairs, with 12,025

Table 1: Quantitative comparison of different methods on the LLVIP, M^{3}FD, and MSRS datasets. The best and second-best performances for each metric are highlighted with Red and Blue backgrounds, respectively.

Dataset	Metric	LRRNet	DDFM	U2Fusion	Text-IF	LENFusion	DIVFusion	ReCoFuse	Ours
LLVIP	EN↑	5.971	6.377	6.033	7.065	7.477	7.292	7.232	**7.529**
	MI↑	1.372	1.774	1.429	1.920	1.724	4.899	2.619	**4.251**
	QABF↑	0.301	0.442	0.372	0.516	0.393	0.396	0.468	**0.526**
	VIF↑	0.267	0.409	0.310	0.490	0.560	0.481	0.498	**0.552**
	QG↑	0.255	0.334	0.294	0.534	0.306	0.433	0.400	**0.540**
M^{3}FD	EN↑	6.619	6.917	6.870	6.833	6.990	7.371	7.090	**7.262**
	MI↑	1.924	1.997	1.825	2.414	1.507	4.059	1.019	**4.548**
	QABF↑	0.432	0.438	0.490	0.557	0.355	0.391	0.567	**0.548**
	VIF↑	0.364	0.396	0.388	0.449	0.644	0.562	0.512	**0.653**
	QG↑	0.288	0.307	0.353	0.379	0.234	0.422	0.505	**0.514**
MSRS	EN↑	6.356	5.769	4.724	7.010	7.193	6.572	7.089	**7.133**
	MI↑	1.921	1.586	1.198	2.319	1.528	0.087	2.410	**2.806**
	QABF↑	0.435	0.420	0.336	0.525	0.265	0.053	0.441	**0.541**
	VIF↑	0.322	0.368	0.250	0.430	0.565	0.030	0.441	**0.573**
	QG↑	0.321	0.305	0.238	0.430	0.174	0.043	0.320	**0.480**

pairs for training and 3,463 pairs for testing. We train PDD-Fusion on the official LLVIP training split and randomly select 300 pairs from its test split for evaluation. To assess cross-dataset generalization, the LLVIP-trained model is directly evaluated on 300 MSRS pairs and 200 M^{3}FD pairs without fine-tuning. All compared methods are evaluated under the same sampling protocol.

Implementation Details. PDD-Fusion is implemented in PyTorch and trained on two NVIDIA L20 GPUs. We use AdamW (Loshchilov & Hutter, 2017) with an initial learning rate of 1×10^{-4} and cosine annealing for 100 epochs. The model is trained without reference fused images using the exposure-adaptive intensity, structural-consistency, and reliability-aware gradient objectives defined in Sec. 3.4.

Evaluation Metrics. We report five widely used image-fusion metrics: Entropy (EN) (Shannon, 1948), Mutual Information (MI), $Q_{AB/F}$ (Xydeas & Petrovic, 2000), Visual Information Fidelity (VIF) (Sheikh & Bovik, 2006), and the gradient-based quality index (QG) (Xydeas & Petrovic, 2000). These metrics evaluate complementary aspects of information content, source-information preservation, and structural quality, but do not directly measure downstream task performance.

4.2 Comparison with State-of-the-art Methods

We compare PDD-Fusion with seven representative fusion methods: LRRNet (Li et al., 2023), DDFM (Zhao et al., 2023b), U2Fusion (Xu et al., 2020), Text-IF (Yi et al., 2024), LENFusion (Chen et al., 2024), DIVFusion (Tang et al., 2023), and ReCoFuse (Zhang et al., 2026). Table 1 reports the quantitative results on LLVIP, M^{3}FD, and MSRS.

On LLVIP, PDD-Fusion ranks first on EN, $Q_{AB/F}$, and QG, and second on MI and VIF, placing among the top two methods on all five metrics. On M^{3}FD, it achieves the best MI, VIF, and QG scores and the second-best EN, while remaining competitive on $Q_{AB/F}$. On MSRS, PDD-Fusion obtains the best results on MI, $Q_{AB/F}$, VIF, and QG, together with the second-best EN. Notably, the results on M^{3}FD and MSRS are obtained using the LLVIP-trained model without fine-tuning. The consistent performance across datasets supports the cross-dataset generalization of the proposed degradation-specific fusion strategy.

Figure 3 provides representative qualitative comparisons on the three datasets. These results complement the quantitative evaluation by illustrating the visual characteristics of the fused outputs under both in-dataset and cross-dataset settings.

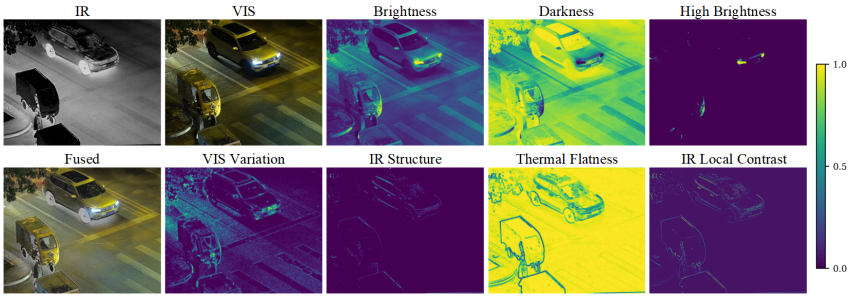

Figure 4: Visualization of the seven physics-guided degradation priors and the fused result.

Table 2: Ablation study of the main components in PDD-Fusion on LLVIP. The best and second-best performances for each metric are highlighted with Red and Blue backgrounds, respectively.

Variant	Physics-Guided Priors	Decoupling Branches	Spatial Prompt	Global Prompt	EN↑	MI↑	QABF↑	VIF↑	QG↑
Baseline	–	–	–	–	7.214	3.652	0.451	0.487	0.421
Module 1	✓	–	–	–	7.356	3.981	0.486	0.521	0.478
Module 2	✓	✓	–	–	7.438	4.126	0.507	0.538	0.512
Module 3	✓	✓	✓	–	7.512	4.236	0.521	0.548	0.533
PDD-Fusion	✓	✓	✓	✓	**7.529**	**4.251**	**0.526**	**0.552**	**0.540**

4.3 Physics-Guided Prior Visualization

Figure 4 visualizes the seven source-derived priors defined in Sec. 3.1, together with the fused result. The visible cues characterize relative exposure and noise-sensitive high-frequency variation, while the infrared cues capture directional structure, thermal flatness, and local contrast. These maps show how observable source-image statistics are converted into spatially structured reliability cues without additional degradation annotations.

4.4 Ablation Study

We conduct a component-wise ablation study on LLVIP to assess the main designs in PDD-Fusion. Starting from the baseline, we progressively introduce physics-guided priors, degradation-decoupling branches, spatial prompt conditioning, and global prompt modulation.

As shown in Table 2, Module 1 evaluates the effect of explicit physics-guided prior modeling. Module 2 further introduces the degradation-decoupling branches without prompt conditioning, while Module 3 adds spatial degradation guidance. The complete PDD-Fusion additionally incorporates global prompt modulation. This progressive design allows the effects of degradation perception, specialized processing, and spatial/global conditioning to be examined separately.

Since the variants are cumulative, this study evaluates the overall design rather than isolating each individual prior or branch. Additional analyses are provided in the Appendix.

5 Conclusion

We presented PDD-Fusion, a physics-guided degradation-decoupled framework for low-light infrared–visible image fusion. Seven source-derived priors are encoded into spatial and global prompts to guide high-frequency reliability modulation, low-frequency exposure-aware fusion, and gated infrared residual compensation. Direct modality paths further preserve complementary source information during reconstruction. Experiments on LLVIP, M^3FD, and MSRS demonstrate strong fusion performance, while direct transfer to M^3FD and MSRS without fine-tuning supports the cross-dataset generalization of the proposed degradation-specific fusion strategy.

AI USE STATEMENT

Generative AI tools were used to assist with code editing, LaTeX formatting, and language polishing. All AI-assisted content and code were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11744–11754. IEEE, 2025.

Jun Chen, Liling Yang, Wei Liu, Xin Tian, and Jiayi Ma. Lenfusion: A joint low-light enhancement and fusion network for nighttime infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 73:1–15, 2024.

Wenhan Yang Jiaying Liu Chen Wei, Wenjing Wang. Deep retinex decomposition for low-light enhancement. In *British Machine Vision Conference*, 2018.

Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1780–1789, 2020.

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

Haiyang Jiang, Huiqin Zhang, Yanduo Zhang, Jiayi Ma, Junjun Jiang, and Huabing Zhou. Phyfusion: Physics-aware infrared and visible image fusion via modality-specific physical priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4993–5002, 2026.

Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-In-One Image Restoration for Unknown Corruption. In *IEEE Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, June 2022.

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.

Hui Li, Xiao-Jun Wu, and Josef Kittler. Rfn-nest: An end-to-end residual fusion network for infrared and visible images. *Information Fusion*, 73:72–86, 2021.

Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):11040–11052, 2023.

Jing Li, Yifan Wang, Jiafeng Yan, Renlong Zhang, and Bin Yang. Mdaif: Robust one-stop multi-degradation-aware image fusion with language-driven semantics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6226–6234, 2026.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5802–5811, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

Vaishnav Potlapalli, Syed Waqas Zamir, Salman Khan, and Fahad Khan. Promptir: Prompting for all-in-one image restoration. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Hamid R Sheikh and Alan C Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006.

Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network. *Information Fusion*, 82: 28–42, 2022a.

Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83: 79–92, 2022b.

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

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.

Wenkuan Xie, Weiguo Pan, Jiancheng Zhang, Songyin Dai, Bingxin Xu, and Cheng Xu. Dcdfusion: A dual-consistency decoupling infrared and visible image fusion algorithm for low-light scenes. *Infrared Physics & Technology*, pp. 106214, 2025.

Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE transactions on pattern analysis and machine intelligence*, 44(1): 502–518, 2020.

Costas S Xydeas and Vladimir Petrovic. Objective image fusion performance measure. *Electronics letters*, 36(4):308–309, 2000.

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

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

Zixiang Zhao, Haowen Bai, 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 (CVPR)*, pp. 5906–5916, June 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*, pp. 8082–8093, 2023b.

A PHYSICS-GUIDED PRIOR ANALYSIS

PDD-Fusion introduces physics-guided priors to characterize degradation-related reliability before multimodal feature interaction. Rather than relying solely on implicitly learned representations, these priors are constructed from observable statistics of infrared and visible images and provide interpretable conditioning information for adaptive fusion.

The seven priors are organized into three groups according to their modality and frequency characteristics: visible exposure reliability, high-frequency reliability, and infrared thermal characteristics. Instead of treating each prior as an independent degradation indicator, the grouped representation captures complementary information that is subsequently associated with degradation-specific processing pathways.

A.1 PHYSICAL MOTIVATION OF PRIOR GROUPING

The grouping strategy follows the observation that different degradation characteristics affect multimodal information in different ways.

First, illumination variation mainly affects the low-frequency intensity distribution of visible images. We therefore use relative brightness, darkness, and high-relative-brightness responses to characterize visible exposure conditions. These cues describe image-relative exposure patterns rather than calibrated physical exposure measurements.

Second, high-frequency responses may contain both useful structures and unreliable fluctuations. The visible variation cue captures local high-frequency activity and may respond to both texture and noise, while the infrared structure cue emphasizes directional structural information. Their combination provides complementary information for high-frequency reliability modeling.

Third, infrared observations also exhibit spatially varying local characteristics. Thermal flatness responds strongly to regions with weak local infrared variation, whereas infrared local contrast captures positive center–surround responses. These cues characterize infrared flatness and local contrast without assuming semantic target saliency.

Based on these observations, the relationship between the proposed physics-guided priors and degradation-specific branches is summarized in Table 3. The three groups provide complementary conditioning information for fusion operations with different modality and frequency characteristics.

Table 3: Summary of the physics-guided prior groups and their corresponding degradation-specific processing branches.

Prior Group	**Exposure**	**High-frequency**	**Infrared thermal**
Derived Cues	Brightness, Darkness, High Brightness	VIS Variation, IR Structure	Thermal Flatness, IR Local Contrast
Representation Meaning	Relative visible exposure	High-frequency variation and structure	Infrared flatness and local contrast
Associated Branch	Low-frequency exposure-aware fusion	High-frequency reliability modulation	Gated infrared residual compensation

A.2 PRIOR-GUIDED DEGRADATION DECOUPLING

The proposed priors are not used to independently enhance individual image components. Instead, they regulate how source information is processed according to its modality and frequency characteristics.

The exposure cues guide low-frequency exposure-aware fusion by adapting the relative contributions of visible and infrared information under different exposure conditions. The high-frequency cues provide information about local variation and structural responses, allowing the corresponding branch to reduce unreliable visible high-frequency contributions while preserving useful structures.

For infrared residual compensation, visible darkness and infrared flatness jointly indicate regions where additional residual processing is activated. The infrared thermal cues further contribute to the learned degradation representation by describing local flatness and contrast characteristics. Importantly, the compensation-need response indicates where additional processing is required rather than directly measuring the usefulness of infrared information.

Consequently, the prior-to-branch design establishes an explicit cue-to-operation correspondence. Rather than handling heterogeneous degradation characteristics through a single shared fusion mechanism, PDD-Fusion assigns information with different modality and frequency properties to specialized processing pathways before adaptive reconstruction.

B Degradation-Decoupling Branch Analysis

The physics-guided priors provide degradation-related conditioning information, while the three specialized branches determine how this information affects multimodal feature processing. Instead of applying a single fusion operation to all features, PDD-Fusion assigns different modality- and frequency-dependent factors to high-frequency reliability modulation, low-frequency exposure-aware fusion, and gated infrared residual compensation.

B.1 High-Frequency Reliability Modulation

Under low-light conditions, visible high-frequency responses may contain both useful structures and unreliable fluctuations. The high-frequency branch therefore combines degradation guidance with high-frequency features to generate a spatial confidence response M_h.

Regions assigned lower confidence receive weaker visible high-frequency contributions, reducing the influence of unreliable fluctuations, while higher-confidence regions preserve more structural information. Since the visible variation cue may respond to both texture and noise, M_h should be interpreted as a learned reliability modulation map rather than an explicit noise mask.

As illustrated in Fig. 5, M_h varies spatially across different structures and imaging conditions, allowing the branch to adapt the contribution of visible high-frequency information instead of uniformly preserving all local responses.

B.2 Low-Frequency Exposure-Aware Fusion

Illumination variation mainly affects the low-frequency intensity distribution of visible images. Directly preserving visible low-frequency information may therefore introduce unreliable intensity responses in poorly exposed regions.

The low-frequency branch predicts modality-specific weighting maps W_{vi} and W_{ir} according to the degradation representation. These weights adapt the relative contributions of visible and infrared low-frequency information across spatial locations.

As shown in Fig. 5, W_{vi} and W_{ir} exhibit spatially varying responses rather than uniformly favoring one modality. This behavior enables the model to adapt low-frequency fusion according to local exposure conditions.

B.3 Gated Infrared Residual Compensation

The infrared branch is designed to provide additional residual processing where visible information is weak and the infrared observation is locally flat. Specifically, the compensation-need map is constructed from visible darkness and infrared flatness as

$$M_t = D_{vi} \odot T_{ir}. \tag{29}$$

A high response therefore indicates a region that is simultaneously dark in the visible image and locally flat in the infrared image.

Importantly, M_t is a compensation-need map rather than an estimator of infrared information usefulness. It gates infrared residual processing in regions requiring additional compensation, while the residual operation still relies on information already present in the infrared observation.

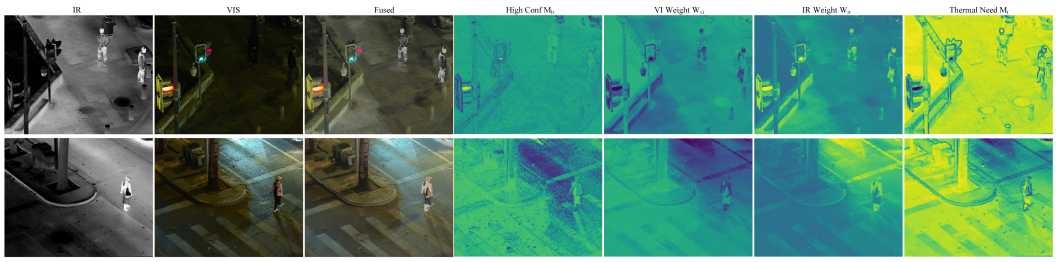

Figure 5: Visualization of degradation-decoupling branch responses. M_h modulates visible high-frequency reliability, W_{vi} and W_{ir} provide spatially adaptive modality weights for low-frequency fusion, and M_t indicates regions where gated infrared residual compensation is activated.

Table 4: Ablation study of different degradation prompt strategies.

Variant	EN↑	MI↑	QABF↑	VIF↑	QG↑
No Prompt	7.438	4.126	0.507	0.538	0.512
Spatial Prompt	7.472	4.178	0.514	0.544	0.524
Global Prompt	7.461	4.162	0.511	0.542	0.521
Spatial + Global Prompt	7.529	4.251	0.526	0.552	0.540

Figure 5 shows that M_t responds strongly in regions satisfying these joint conditions, complementing the spatial behavior of the high- and low-frequency branches.

B.4 Visualization of Branch Responses

Figure 5 provides representative responses of the three degradation-decoupling branches. The high-frequency confidence map M_h regulates visible detail preservation, the modality weights W_{vi} and W_{ir} control low-frequency mixing, and the compensation-need map M_t gates infrared residual processing.

The distinct spatial patterns of these responses illustrate the proposed cue-to-operation correspondence: different degradation-related factors produce different processing responses rather than being handled through a single shared interaction mechanism.

C Additional Ablation Studies

Beyond the cumulative component analysis in the main paper, we further investigate two design choices in PDD-Fusion: the combination of spatial and global degradation prompts and the FiLM-based global modulation mechanism. These experiments provide a more detailed analysis of how degradation conditioning contributes to fusion performance.

C.1 Ablation Study on Prompt Guidance Strategies

The degradation representation module produces two complementary forms of guidance. Spatial prompts preserve location-dependent degradation information, whereas global prompts summarize image-level degradation conditions for adaptive feature modulation.

We compare four variants: removing both prompts, using only the spatial prompt, using only the global prompt, and using both prompts. The results are reported in Table 4.

Both spatial-only and global-only variants outperform the no-prompt setting on all five metrics, indicating that each form of degradation conditioning provides useful information. The spatial prompt performs slightly better than the global prompt when used independently, while their combination achieves the best results across all metrics. This suggests that local degradation patterns and image-level degradation conditions provide complementary guidance for the fusion process.

C.2 Ablation Study on Feature Modulation Mechanism

PDD-Fusion employs FiLM-based modulation to incorporate the global degradation representation into branch features. To examine its effect, we compare the complete model with a variant in which all FiLM operations are removed while the remaining architecture is unchanged.

Table 5: Ablation study of FiLM-based feature modulation.

Variant	EN↑	MI↑	QABF↑	VIF↑	QG↑
w/o FiLM	7.468	4.129	0.518	0.612	0.516
Full FiLM	7.529	4.251	0.526	0.552	0.540

As shown in Table 5, FiLM modulation improves EN, MI, QABF, and QG, whereas the variant without FiLM obtains a higher VIF. Thus, the effect of FiLM is not uniform across all metrics. Nevertheless, the full model achieves stronger performance on four of the five reported criteria, supporting the use of global degradation conditioning for adaptive feature regulation.

C.3 Discussion

The additional ablations show that the proposed degradation conditioning is supported by both local and global information. Spatial prompts provide location-dependent guidance, while global prompts summarize the overall imaging condition, and their combination consistently gives the strongest results.

FiLM further introduces image-level modulation into the branch features. Although its effect involves a trade-off in VIF, it improves the remaining four metrics. Together with the degradation-decoupling branches, these components form a hierarchical conditioning pipeline from degradation representation to adaptive feature processing.

D Computational Complexity Analysis

We compare the parameters, FLOPs, and inference time of PDD-Fusion with representative IVIF methods under the same hardware and input resolution.

Table 6: Computational complexity comparison. Params, FLOPs, and Time denote trainable parameters, computational cost, and inference latency, respectively.

Method	**Params (M)**	**FLOPs (G)**	**Time (ms)**
LRRNet	**0.21**	**40.05**	51.52
DDFM	552.81	5220.50	8355.95
U2Fusion	0.66	43.17	**4.99**
Text-IF	215.12	325.55	80.77
LENFusion	0.66	61.10	26.30
DIVFusion	4.40	180.64	271.94
ReCoFuse	155.47	254.13	2959.87
PDD-Fusion	0.90	103.30	<u>9.68</u>

As shown in Table 6, PDD-Fusion contains only 0.90M trainable parameters and requires 103.30G FLOPs, substantially less than large-scale methods such as DDFM, Text-IF, and ReCoFuse. Its inference time of 9.68 ms is also lower than most compared methods. These results show that the additional degradation-aware processing maintains compact parameterization and low inference latency without relying on a large increase in model capacity.