RETHINKING INFRARED-VISIBLE IMAGE FUSION VIA SEMANTIC DIAGNOSIS AND REWARD-DRIVEN OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Infrared and visible (IR-VIS) image fusion aims to integrate thermal target saliency from IR images and textural details from VIS images into a single fused image. Existing methods mainly rely on low-level visual fidelity objectives, emphasizing image sharpness and structural consistency while overlooking semantic preservation and task-relevant cues. Moreover, the lack of ground-truth fusion labels and scene-dependent trade-offs among thermal saliency, texture details, structural consistency, and artifact suppression make fixed loss functions inadequate for adaptive optimization. To address these limitations, we propose a semantic-diagnosis- and reward-driven framework that combines high-level semantic guidance with candidate-based reward optimization. Specifically, we introduce a frozen DINOv3-ConvNeXt to provide hierarchical visual priors and global visual representations, together with a lightweight adaptation branch for local texture and detail enhancement. Building on these representations, we leverage a frozen Qwen3-VL-Embedding to perform coarse-to-fine semantic diagnosis on intermediate fusion states and progressively guide their refinement. To further enable adaptive optimization without fusion labels, we generate multiple fusion candidates through perturbation sampling and evaluate their relative quality using multidimensional rewards. We then employ an advantage-weighted objective to optimize the candidate generation policy, thereby learning fusion policies through candidate comparison and reward optimization. Experimental results demonstrate that our method consistently outperforms existing approaches in both quantitative metrics and visual quality across multiple public benchmarks.

1 INTRODUCTION

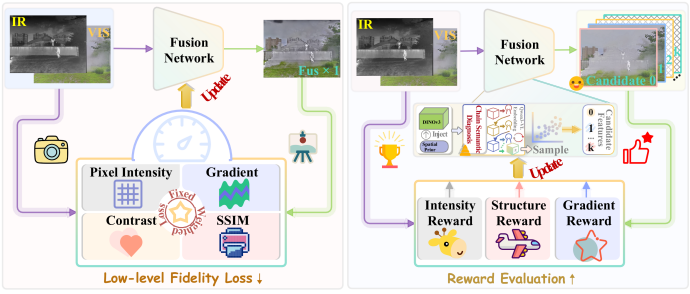
Infrared and visible (IR-VIS) image fusion aims to integrate thermal target saliency from infrared images with textural details from visible images. It has been widely applied (Liu et al., 2025a) to scene perception, UAV target reconnaissance, medical image-aided diagnosis, and target detection and semantic segmentation. In recent years, deep learning-based methods have shifted from handcrafted fusion rules to end-to-end learning, substantially improving fusion performance. Studies have also explored joint alignment and fusion for unregistered IR-VIS images (Li et al., 2025; Chen et al., 2026). However, most existing methods still rely primarily on low-level visual fidelity objectives. They typically use predefined constraints on pixel intensity, gradients, contrast, or structural similarity to supervise a single fused output, as shown in Figure 1(a). Although these constraints improve sharpness and structural consistency, they do not explicitly ensure the preservation of target semantics, scene structure, and task-relevant cues (Yang et al., 2026) throughout the fusion process.

This low-level fidelity-driven fusion paradigm leads to two direct consequences. First, the model may overemphasize gradient enhancement or pixel intensity consistency, resulting in spurious textures, local artifacts, or unnatural structures in fused images. Second, even when the fused image achieves high scores on objective metrics, thermal targets may still be weakened, target boundaries may become blurred, and key semantic regions may be insufficiently preserved. These issues can further degrade downstream tasks such as detection, recognition, and scene understanding. More importantly, IR-VIS image fusion typically lacks ground-truth fusion labels. Meanwhile, different scenes require different trade-offs among thermal saliency, texture details and structural consistency.

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.

Therefore, a fixed loss function cannot adequately capture these dynamic quality trade-offs or provide adaptive guidance for different fusion states.

Based on the above analysis, we argue that IR-VIS image fusion should not be treated merely as a fixed error-minimization problem. Instead, it should be formulated as a layer-wise diagnosable and dynamically optimizable generation process. To this end, the model should support three key capabilities. First, under limited fusion training data, it should leverage stable general visual priors to strengthen the representation of global structures and multi-scale visual information. Second, it should monitor intermediate fusion states and identify semantic degradation, such as target weakening, structural loss, and spurious texture enhancement. Third, in the absence of ground-truth fusion labels, it should learn better fusion policies through quality comparisons among multiple candidates, rather than statically constraining a single output with a fixed weighted loss.



(a) Conventional Fusion with Fixed Low-level Losses (b) Reward-driven Multi-candidate Fusion Exploration

Figure 1: Comparison between the traditional fusion paradigm and our framework.

To address these limitations, we reformulate IR-VIS image fusion through semantic diagnosis and reward-driven optimization, as shown in Figure 1(b). Rather than treating fusion as a single-step low-level reconstruction problem, our framework models it as a progressive generation process involving visual prior construction, intermediate-state diagnosis, and candidate-based reward optimization. Specifically, we employ a frozen DINOv3-ConvNeXt (Siméoni et al., 2026) to extract stable multi-scale visual priors and global structural representations. In parallel, lightweight adaptation branches enhance local textures, edges, and high-frequency details. Together, they provide complementary global and local representations for subsequent fusion.

Based on these representations, we further incorporate a frozen Qwen3-VL-Embedding (Li et al., 2026a) for coarse-to-fine semantic diagnosis of intermediate fusion states. Instead of evaluating only the final output, it identifies semantic degradation at different stages and converts the feedback into cross-layer channel modulation signals for progressive refinement. After semantic diagnosis and modulation, multiple candidates are generated in the final feature space through feature perturbation. Their relative quality is evaluated using multidimensional rewards that consider thermal saliency, texture preservation, structural consistency, and artifact suppression. We optimize the candidate generation policy via a GRPO-inspired differentiable advantage objective (Williams, 1992; Shao et al., 2024), encouraging higher-quality results through relative comparison rather than a single fixed loss. The main contributions of this paper are as follows:

- **We propose a foundation-model-guided visual prior construction and progressive fusion scheme.** To overcome the limited representation capability of conventional fusion networks, we introduce a frozen DINOv3-ConvNeXt to provide stable multi-scale visual priors and global structural representations. We further design lightweight adaptation branches to enhance local textures, edges, and high-frequency details. This design enables progressive integration of global semantics and fine-grained visual information.
- **We develop a chained semantic diagnosis and feedback modulation mechanism for intermediate fusion states.** Unlike existing methods that mainly supervise the final fused output, our method explicitly diagnoses intermediate fusion states in a coarse-to-fine manner using a frozen Qwen3-VL-Embedding. We further transform the diagnostic feedback into cross-layer channel modulation signals, allowing the network to progressively correct semantic degradation during fusion.
- **We present a candidate-comparison-based reward optimization strategy for fusion learning.** Instead of optimizing a single fused output with a fixed weighted loss, we generate multiple candidates through feature perturbation and evaluate their relative quality using multidimensional rewards. We then optimize the candidate generation policy with advantage weighting, enabling the model to learn fusion preferences through candidate comparison and reward optimization.

2 Related Work

2.1 Representation and Semantic Guidance for IR-VIS Image Fusion

In IR-VIS image fusion, existing deep fusion methods commonly rely on encoder-decoder architectures, feature refinement, attention mechanisms, and cross-modal interaction. For example, U2Fusion (Xu et al., 2022), DenseFuse (Li & Wu, 2019), RFN-Nest (Li et al., 2021b), DATFuse (Tang et al., 2023), SwinFusion (Ma et al., 2022), and HitFusion (Chen et al., 2024) improve feature extraction and cross-modal fusion through hierarchical representations and interaction mechanisms. CDDFuse (Zhao et al., 2023), SMLNet (Kang et al., 2025), and DRFusion (Shi et al., 2026) further explore feature decomposition, manifold modeling, and generative learning, respectively. However, most of these methods are still dominated by low-level fidelity objectives, such as pixel, gradient, and structural constraints. Such objectives do not explicitly ensure the preservation of target semantics, scene structures, or task-relevant cues.

With the development of foundation vision models, recent studies have explored pre-trained priors to improve fusion representations and generalization. For example, UniFusion (Li et al., 2026c) employs DINOv3 to construct a modality-consistent feature space. However, generic visual priors may not sufficiently capture local textures, edges, and high-frequency details required by IR-VIS fusion. Our method therefore combines global visual priors with lightweight fine-grained adaptation to preserve both global structures and local details. Semantic information has also been introduced to improve fusion quality and downstream performance. Text-IF (Yi et al., 2024), TextFusion (Cheng et al., 2025), OmniFuse (Zhang et al., 2025), MTG-Fusion (Wang et al., 2025b), ControlFusion (Tang et al., 2025), and AWM-Fuse (Li et al., 2026b) guide fusion using text prompts, task instructions, or vision-language models. However, semantic information in these methods is mainly used as an input condition, external cue, or final-output guidance. It therefore provides limited capability to identify semantic degradation, such as target weakening, structural loss, or spurious texture generation, during intermediate fusion. In contrast, our method performs semantic diagnosis directly on intermediate fusion states and converts the diagnostic feedback into modulation signals for progressive correction.

2.2 Fusion Optimization Paradigm and Reward-Driven Learning

IR-VIS image fusion typically lacks ground-truth fusion labels, while different scenes require different trade-offs among thermal saliency, texture details, structural consistency, and artifact suppression. This makes fixed loss functions insufficient for adaptive fusion optimization. Most existing methods rely on predefined objectives, such as pixel consistency, structural similarity, gradient preservation, perceptual loss, or task loss. These objectives are balanced using fixed weights, dynamic weighting, or loss-function search. For example, RFN-Nest (Li et al., 2021b) and FusionGAN (Ma et al., 2019) optimize fusion quality through combinations of pixel, gradient, structural, or adversarial losses. TIMFusion (Liu et al., 2024), DCEvo (Liu et al., 2025b), LDFusion (Wang et al., 2025a), and TDFusion (Bai et al., 2025) further explore semantic constraints, dynamic weighting, or loss-function search. However, these methods still optimize a single fused output under predefined objectives and do not explicitly compare the relative quality of multiple candidates.

Recent studies have explored reward-driven learning for fusion without ground-truth labels. For example, *Bridging Human Evaluation to IR-VIS Image Fusion* (Liu et al., 2026) learns fusion preferences from human feedback and optimizes the network with GRPO. However, it requires external preference data and an additional reward model. In contrast, our method directly compares multiple candidates generated through feature perturbation, evaluates them with multidimensional rewards, and optimizes the candidate generation policy using advantage weighting. This enables candidate-comparison-driven reward optimization without relying on human preference supervision.

3 Proposed Method

3.1 Overview

We propose a unified IR-VIS image fusion framework, as illustrated in Figure 2. It consists of three modules: DINOv3-based Hierarchical Feature Extraction and Progressive Fusion (**DHFE-PF**), Chained Semantic Diagnosis and Progressive Regulation (**CSD-PR**), and Candidate Generation and

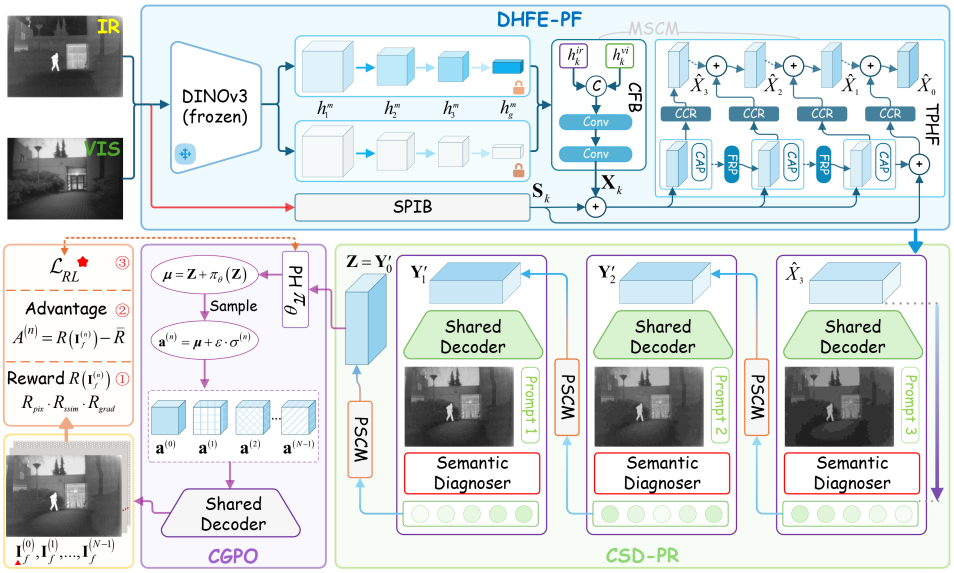


Figure 2: Overall framework of the proposed method. The model combines DINOv3-based multi-scale representation learning, Qwen3-VL-Embedding-based chained semantic diagnosis, and advantage-weighted candidate generation and policy optimization for IR-VIS image fusion.

Policy Optimization (**CGPO**). DHFE-PF combines frozen DINOv3 visual priors with lightweight local adaptation to construct multi-scale fusion representations. CSD-PR diagnoses intermediate fusion states using a frozen Qwen3-VL-Embedding and progressively converts semantic feedback into feature modulation. CGPO generates multiple fusion candidates and optimizes their sampling policy through multidimensional rewards and advantage weighting.

3.2 Feature Extraction and Progressive Fusion

3.2.1 Hierarchical Visual Prior Extraction

IR-VIS image fusion requires both global structural representation and fine-grained detail preservation. However, fusion models are often trained on relatively limited datasets, which may restrict their representation capability and cross-scene generalization. We therefore employ a pre-trained and frozen DINOv3-ConvNeXt as the feature encoder Φ to extract stable hierarchical visual priors. Let $\mathbf{I}_m$, $m \in \{ir, vis\}$, denote the input image of modality m. The encoder extracts hierarchical features as $\{\mathbf{h}_1^m, \mathbf{h}_2^m, \mathbf{h}_3^m, \mathbf{h}_g^m\} = \Phi(\mathbf{I}_m)$, where $\mathbf{h}_1^m$, $\mathbf{h}_2^m$, and $\mathbf{h}_3^m$ denote multi-scale visual features, and $\mathbf{h}_g^m = \{\mathbf{h}_g^m[\tau]\}_{\tau=1}^T$ denotes a global feature sequence of length T.

These hierarchical features provide stable visual priors, but fine-grained textures, edges, and high-frequency details may be gradually weakened by progressive downsampling during hierarchical feature extraction. To compensate for this information loss, we further introduce a Spatial Prior Injection Branch (SPIB), which extracts local information directly from the input images and generates multi-scale adaptation features $\mathbf{S}_k = \psi_k\big(\mathcal{C}_{\text{prior}}(\text{Cat}[\mathbf{I}_{ir}, \mathbf{I}_{vis}])\big), \quad k \in \{0, 1, 2, 3\}$, where $\mathcal{C}_{\text{prior}}$ is the prior extractor in SPIB, which extracts a spatial prior from the concatenated IR-VIS inputs, and ψ_k is the scale-specific mapping function that projects this prior to the k-th feature level.

3.2.2 Multi-source Collaborative Fusion

Based on the hierarchical features and spatial priors, we construct a Multi-source Collaborative Fusion Module (MSCM) to jointly integrate cross-modal, cross-scale, and local information. At each scale k, the IR and VIS features are first fused by a Cross-modal Fusion Block (CFB): $\mathbf{X}_k = \mathcal{C}_{\text{fuse}}^k\big(\text{Cat}[\mathbf{h}_k^{ir}, \mathbf{h}_k^{vis}]\big)$, where $\mathcal{C}_{\text{fuse}}^k$ denotes the cross-modal fusion block at the k-th scale.

For $k < 3$, the higher-level fused feature is interpolated to the current resolution and injected into the current representation through a Channel Alignment Projection (CAP) $\mathcal{P}_k$:

$$\mathbf{X}_k \leftarrow \mathbf{X}_k + \mathcal{P}_k\big(\text{Interp}(\mathbf{X}_{k+1}^{raw}, \text{size}(\mathbf{S}_k))\big). \tag{1}$$

Here, $\mathrm{Interp}(\cdot)$ denotes bilinear interpolation. For $k = 3$, cross-scale injection is omitted because no higher-level fused feature is available. The local adaptation feature $\mathbf{S}_k$ is then added as $\mathbf{X}_k \leftarrow \mathbf{X}_k + \mathbf{S}_k$, followed by a Feature Refinement Projection (FRP) $\mathcal{C}^k_{\mathrm{refine}}$, yielding $\mathbf{X}^{raw}_k = \mathcal{C}^k_{\mathrm{refine}}(\mathbf{X}_k)$.

Within MSCM, we further construct a Top-down Progressive Hierarchical Fusion (TPHF) pathway to aggregate refined representations across scales. Unlike the preceding cross-scale injection, which operates at native feature resolutions, TPHF aggregates refined features after mapping them to a common spatial resolution. This allows high-level structural information to guide lower-level texture and detail refinement. Let $[H, W]$ denote the input resolution. TPHF employs a Channel Compression and Refinement (CCR) module $\mathcal{R}_k$ for channel alignment and feature refinement:

$$\hat{\mathbf{X}}_k = \mathcal{R}_k\big(\mathrm{Interp}(\mathbf{X}^{raw}_k, [H, W])\big) + \hat{\mathbf{X}}_{k+1}, \quad k = 3, 2, 1, \tag{2}$$

where $\hat{\mathbf{X}}_4 = \mathbf{0}$. At the input scale, the full-resolution adaptation feature $\mathbf{S}_0$ is further incorporated:

$$\hat{\mathbf{X}}_0 = \mathcal{R}_0\big(\mathrm{Interp}(\mathbf{X}^{raw}_1, [H, W]) + \mathbf{S}_0\big) + \hat{\mathbf{X}}_1. \tag{3}$$

This produces the multi-level fusion features $\mathcal{D} = \{\hat{\mathbf{X}}_0, \hat{\mathbf{X}}_1, \hat{\mathbf{X}}_2, \hat{\mathbf{X}}_3\}$, all at the input resolution with the same channel dimension.

3.3 Chained Semantic Diagnosis and Progressive Regulation

Existing fusion methods often use large-scale models for final-result evaluation, text prompting, or external semantic guidance. However, they provide limited support for refining intermediate fusion states. We instead embed a frozen Qwen3-VL-Embedding into the fusion pathway for chained diagnosis. Intermediate features are first reconstructed into stage-wise fusion images. These images are then evaluated with stage-specific prompts. The resulting semantic feedback is progressively propagated to finer stages for feature regulation.

3.3.1 Intermediate Fusion Reconstruction and Semantic Feedback

Given the fusion features $\mathcal{D}$, we initialize the stage features as $\mathbf{Y}_k = \hat{\mathbf{X}}_k$. To distinguish the original stage feature from its semantically modulated state, we denote the current diagnostic feature by $\mathbf{Y}^c_k$. At the highest stage, $\mathbf{Y}^c_3 = \mathbf{Y}_3$. At subsequent stages, $\mathbf{Y}^c_k$ is obtained by modulating $\mathbf{Y}_k$ with semantic feedback from the preceding higher-level stage. At stage k, $\mathbf{Y}^c_k$ is concatenated with the original bimodal inputs $\mathbf{I}_{ir}$ and $\mathbf{I}_{vis}$ along the channel dimension. A shared lightweight decoder $g(\cdot)$ then reconstructs an intermediate fusion image: $\tilde{\mathbf{I}}_k = g\big(\mathrm{Cat}[\mathbf{Y}^c_k, \mathbf{I}_{ir}, \mathbf{I}_{vis}]\big)$. This reconstruction converts the current feature-level fusion state into an image-level representation that can be directly processed by Qwen3-VL-Embedding.

Based on these intermediate fusion images, we construct a chained diagnosis-and-regulation process: $\mathrm{Stage}_3 \rightarrow \mathrm{Stage}_2 \rightarrow \mathrm{Stage}_1 \rightarrow \mathrm{Stage}_0$. Semantic diagnosis is explicitly performed at $k = 3, 2, 1$, while Stage_0 only receives the semantic feedback propagated from Stage_1. At each diagnostic stage, the frozen Qwen3-VL-Embedding Ψ takes the intermediate fusion image $\tilde{\mathbf{I}}_k$, a fixed instruction, and a stage-specific prompt prompt_k as inputs. It outputs a 2048-dimensional semantic vector $\mathbf{l}_k$ that characterizes the current fusion state and its potential deficiencies: $\mathbf{l}_k = \Psi\big(\tilde{\mathbf{I}}_k, \mathrm{Instruction}, \mathrm{Prompt}_k\big)$. The prompts are tailored to the semantic focus of each stage. The high-level prompt ($k = 3$) emphasizes salient objects and global structural consistency. The mid-level prompt ($k = 2$) focuses on object boundaries, regional contrast, and structural continuity. The low-level prompt ($k = 1$) targets textural details, local edges, and artifact suppression. This stage-specific design aligns semantic diagnosis with the hierarchical fusion process and avoids applying identical criteria to different feature levels.

3.3.2 Visual-Semantic Alignment and Chained Modulation

Because Qwen3-VL-Embedding and the fusion network operate in different representation spaces, we first align the semantic feedback with the global visual priors before feature modulation. A joint global visual representation $\mathbf{v}_g = \frac{1}{2T}\sum_{\tau=1}^{T}\big(\mathbf{h}^{ir}_g[\tau] + \mathbf{h}^{vis}_g[\tau]\big)$ is obtained from the IR and VIS global feature sequences. The semantic vector is first projected into the visual representation space: $\hat{\mathbf{l}}_k = \mathbf{W}_q\mathbf{l}_k$. It is then combined with the global visual representation and mapped to the

channel dimension of the next finer-scale feature: $\mathbf{f}_k = \mathbf{W}_{proj}^{(k)}\big(\hat{\mathbf{l}}_k + \mathbf{v}_g\big)$, where $\mathbf{W}_{proj}^{(k)}$ denotes the stage-specific projection for aligning the semantic feedback with $\mathbf{Y}_{k-1}$.

Based on $\mathbf{f}_k$, we generate a channel-wise modulation signal using a centered gate: $\boldsymbol{\gamma}_k = \sigma(\mathbf{f}_k) - 0.5$, which allows both channel enhancement and suppression. Rather than modifying the current stage, semantic feedback from stage k regulates the next finer-scale feature through Progressive Semantic Channel Modulation (PSCM):

$$\mathbf{Y}'_{k-1} = \mathbf{Y}_{k-1} + \mathbf{Y}_{k-1} \odot \boldsymbol{\gamma}_k. \tag{4}$$

The modulated feature then serves as the current state for the next semantic diagnosis: $\mathbf{Y}^c_{k-1} = \mathbf{Y}'_{k-1}$. Accordingly, the chained modulation follows $\mathbf{Y}_3 \to \mathbf{Y}'_2 \to \mathbf{Y}'_1 \to \mathbf{Y}'_0$. This process progressively transfers semantic feedback from high-level structural information to low-level texture and detail refinement. The final semantically refined feature is denoted as $\mathbf{Z} = \mathbf{Y}'_0$.

3.4 Candidate Generation and Policy Optimization

3.4.1 Candidate Policy and Sampling

Fixed weighted losses optimize a single fused output and provide no explicit comparison among alternative fusion states. To address this limitation, CGPO constructs a local candidate distribution around the semantically refined feature $\mathbf{Z}$, allowing the model to explore nearby fusion states and compare their relative quality. To parameterize this distribution, we introduce a lightweight policy head π_θ that predicts a residual offset. The policy mean is defined as $\boldsymbol{\mu} = \mathbf{Z} + \pi_\theta(\mathbf{Z})$. We then define an isotropic Gaussian candidate distribution: $\pi_\theta(\mathbf{a} \mid \mathbf{Z}) = \mathcal{N}\big(\boldsymbol{\mu}, \sigma^2\mathbf{I}\big)$. One reference candidate and $N-1$ perturbed candidates are generated as

$$\mathbf{a}^{(0)} = \boldsymbol{\mu}, \mathbf{a}^{(n)} = \boldsymbol{\mu} + \sigma\boldsymbol{\varepsilon}^{(n)}, \quad \boldsymbol{\varepsilon}^{(n)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad n = 1, \ldots, N-1. \tag{5}$$

Candidate sampling is used only during training. During inference, the reference candidate $\mathbf{a}^{(0)}$ is directly decoded. The decoder $g(\cdot)$ is shared with the intermediate reconstruction process in CSD-PR. Each candidate is reconstructed as

$$\mathbf{I}_f^{(n)} = g\Big(\mathrm{Cat}\big[\mathbf{a}^{(n)}, \mathbf{I}_{ir}, \mathbf{I}_{vis}\big]\Big), \quad n = 0, 1, \ldots, N-1. \tag{6}$$

These reconstructed candidates provide the basis for subsequent reward evaluation and relative quality comparison.

3.4.2 Multidimensional Reward

To compare candidate quality from complementary perspectives, we construct a composite reward based on pixel response consistency, structural similarity, and gradient detail preservation:

$$R(\mathbf{I}_f^{(n)}) = R_{pix}(\mathbf{I}_f^{(n)}) \cdot R_{ssim}(\mathbf{I}_f^{(n)}) \cdot R_{grad}(\mathbf{I}_f^{(n)}). \tag{7}$$

The multiplicative form encourages consistently high quality across all three dimensions. Since the fused output is single-channel, the VIS image is converted to its luminance component for reward computation, denoted as $\mathbf{I}_{vis}^Y$. The pixel response reward is defined as

$$R_{pix}(\mathbf{I}_f^{(n)}) = \frac{1}{1 + \frac{1}{HW}\left\|\mathbf{I}_f^{(n)} - \mathbf{I}_{ref}\right\|_1}, \tag{8}$$

where $\mathbf{I}_{ref} = \max(\mathbf{I}_{ir}, \mathbf{I}_{vis}^Y)$. The structural reward is defined as

$$R_{ssim}(\mathbf{I}_f^{(n)}) = \mathrm{Clip}\left(\frac{1}{2}\left[\mathrm{SSIM}(\mathbf{I}_f^{(n)}, \mathbf{I}_{ir}) + \mathrm{SSIM}(\mathbf{I}_f^{(n)}, \mathbf{I}_{vis}^Y)\right], \varepsilon, 1\right). \tag{9}$$

For gradient preservation, we define $g_{ref} = \max(g_{ir}, g_{vis})$, where g_{ir}, g_{vis}, and $g_f^{(n)}$ denote the Sobel gradient magnitudes of $\mathbf{I}_{ir}$, $\mathbf{I}_{vis}^Y$, and $\mathbf{I}_f^{(n)}$, respectively. The gradient reward is

$$R_{grad}(\mathbf{I}_f^{(n)}) = \frac{1}{1 + \frac{1}{HW}\left\|g_f^{(n)} - g_{ref}\right\|_1}. \tag{10}$$



Figure 3: Qualitative comparison of the proposed method with state-of-the-art IR-VIS image fusion methods on the MSRS, M3FD, and FMB datasets.

3.4.3 Advantage Estimation and Policy Optimization

The above rewards measure the absolute quality of individual candidates. However, our policy optimization uses a relative signal indicating whether a candidate performs better or worse than the alternatives. We therefore convert the candidate rewards into relative advantages using a group-mean baseline. For each candidate $n = 0, \ldots, N-1$,

$$\bar{R} = \frac{1}{N} \sum_{j=0}^{N-1} R(\mathbf{I}_f^{(j)}), \qquad A^{(n)} = R(\mathbf{I}_f^{(n)}) - \bar{R}. \tag{11}$$

For each candidate, its log-probability under the Gaussian policy is computed and normalized over the feature dimensions. The detailed formulation is provided in the Appendix A.3.2. We then employ relative advantages to optimize the candidate generation policy:

$$\mathcal{L}_{RL} = -\frac{1}{N} \sum_{n=0}^{N-1} A^{(n)} \log \pi_\theta\big(\mathbf{a}^{(n)} \mid \mathbf{Z}\big), \tag{12}$$

where $\log \pi_\theta$ denotes the log-probability averaged over the feature dimensions. Minimizing $\mathcal{L}_{\mathrm{RL}}$ updates the policy through differentiable candidate rewards and the group-mean baseline, with further details provided in the Appendix A.3.

4 Experiments

4.1 Datasets and Evaluation Metrics

We train the proposed model only on the MSRS training set (Tang et al., 2022), which contains 1,083 registered IR-VIS image pairs. Evaluation is conducted on MSRS, M3FD (Liu et al., 2022), and FMB (Liu et al., 2023), with 361, 300, and 280 test pairs, respectively. These datasets cover diverse scenes and illumination conditions, enabling evaluation of fusion quality and cross-dataset generalization. We adopt six commonly used metrics: Mutual Information (Q_{MI}) (Qu et al., 2002), Gradient-based Fusion Metric ($Q_{\mathrm{AB/F}}$) (Xydeas & Petrovic, 2000), Visual Information Fidelity (Q_{VIF}) (Sheikh & Bovik, 2006), Feature Contrast-based Metric (Q_{CV}) (Chen & Varshney, 2007), Local Structural Similarity-based Metric (Q_{Y}) (Yang et al., 2008), and Structural Similarity Index (Q_{SSIM}) (Wang et al., 2004). Higher values are better for all metrics except Q_{CV}, where lower values are preferred.

4.2 Comparison with SOTA Methods

We compare the proposed method with eight representative fusion approaches: MetaLFusion (Li et al., 2021a), DATFuse (Tang et al., 2023), TIMFusion (Liu et al., 2024), LUT-Fuse (Yi et al., 2025), SMLNet (Kang et al., 2025), FreeFusion (Zhao et al., 2025), DPOFusion (Su et al., 2026), and DRFusion (Shi et al., 2026).

Table 1: Average quantitative results of different fusion methods on the MSRS, M3FD, and FMB datasets. The best and second-best results are highlighted in **pink** and **blue**, respectively.

Methods	$Q_{\mathrm{MI}}\uparrow$	$Q_{\mathrm{AB/F}}\uparrow$	$Q_{\mathrm{VIF}}\uparrow$	$Q_{\mathrm{CV}}\downarrow$	$Q_{\mathrm{Y}}\uparrow$	$Q_{\mathrm{SSIM}}\uparrow$
MetaLFusion Li et al. (2021a)	3.3804	0.3926	0.6619	569.1025	0.8220	1.3730
DATFuse Tang et al. (2023)	3.9457	0.5629	0.8117	391.1530	0.8841	1.3480
TIMFusion Liu et al. (2024)	3.7357	0.5401	0.7634	514.0484	0.8985	1.3578
LUT-Fuse Yi et al. (2025)	3.8308	0.5397	0.8797	399.7892	0.8906	1.4118
SMLNet Kang et al. (2025)	3.0207	0.5496	0.8141	426.7363	0.9008	1.3974
FreeFusion Zhao et al. (2025)	2.0876	0.4819	0.6679	728.9932	0.7990	1.1697
DPOFusion Su et al. (2026)	1.7741	0.2074	0.6916	809.8966	0.6044	1.1144
DRFusion Shi et al. (2026)	3.4305	0.4414	0.7991	381.5354	0.9042	1.3938
Ours	**4.2216**	**0.6527**	**0.9684**	**336.4684**	**0.9468**	**1.4243**

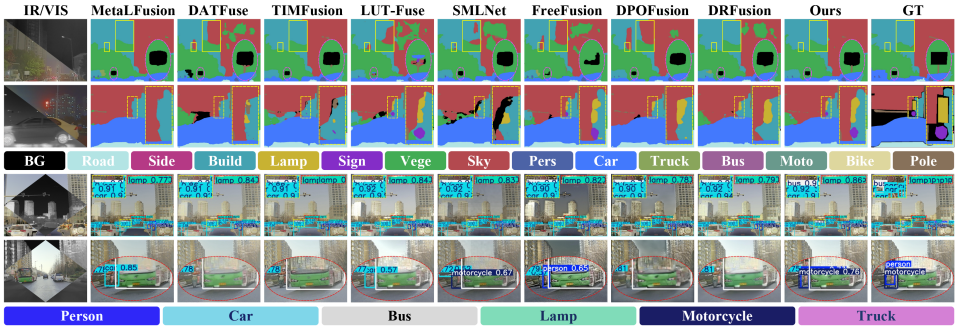


Figure 4: Qualitative comparison of different fusion methods on semantic segmentation and object detection tasks.

4.2.1 Qualitative Comparison

Figure 3 shows qualitative comparisons on the MSRS, M3FD, and FMB datasets. The methods exhibit clear differences in thermal target preservation, visible texture recovery, and overall visual quality. DRFusion produces sharp results but may introduce over-enhancement, distortion, or color shifts in challenging scenes. FreeFusion often provides relatively weak target and structure preservation. LUT-Fuse and DPOFusion perform well overall, but LUT-Fuse may lose visible details, while DPOFusion can produce insufficient brightness under low-light or complex-background conditions. SMLNet occasionally introduces blocking artifacts, whereas TIMFusion, DATFuse, and MetaLFusion may weaken salient infrared targets in some scenes. In contrast, our method consistently preserves thermal targets while maintaining visible structural details across all three datasets, demonstrating better visual robustness under diverse conditions.

4.2.2 Quantitative Comparison

Table 1 reports the average quantitative results over MSRS, M3FD, and FMB, while per-dataset results are provided Table 4 in Appendix A.6. Our method achieves the best average performance on all six metrics. Specifically, Q_{MI} reaches 4.2216, improving over the second-best DATFuse by approximately 7.0%, while $Q_{\mathrm{AB/F}}$ reaches 0.6527 with an improvement of approximately 16.0%. For Q_{VIF}, our method achieves 0.9684, outperforming LUT-Fuse by approximately 10.1%. For Q_{CV}, where lower values are better, our method obtains 336.4684, corresponding to an 11.8% reduction relative to DRFusion. In addition, Q_{Y} and Q_{SSIM} reach 0.9468 and 1.4243, improving over the corresponding second-best results by approximately 4.7% and 0.9%, respectively. These results demonstrate consistent advantages in information preservation, edge fidelity, visual quality, and structural consistency across the three datasets.

4.3 Downstream Task Evaluation

To assess the utility of fused images for downstream tasks, we evaluate semantic segmentation and object detection. For semantic segmentation, SegFormer-B5 (Xie et al., 2021) is applied to FMB, with performance measured by mean Intersection over Union (mIoU). For object detection, YOLOv8s (Jocher et al., 2023) is evaluated on the M3FD dataset using $\mathrm{mAP}_{50:95}$. Qualitative and quantitative results are shown in Fig. 4 and Table 2, respectively. Our method achieves the

Table 3: Ablation results of the core modules and key designs on the MSRS dataset. ✓ and – indicate whether the corresponding module is enabled or disabled. The best and second-best results are highlighted in pink and blue, respectively.

Setting	DHFE-PF		Optimization		Quantitative Metrics					
	DINOv3	SPIB	CSD-PR Stages	CGPO	$Q_{\mathrm{MI}}\uparrow$	$Q_{\mathrm{AB/F}}\uparrow$	$Q_{\mathrm{VIF}}\uparrow$	$Q_{\mathrm{CV}}\downarrow$	$Q_{\mathrm{Y}}\uparrow$	$Q_{\mathrm{SSIM}}\uparrow$
S1	–	✓	0	–	3.9092	0.6374	1.0004	236.5207	0.9719	**1.4278**
S2	✓	–	0	–	4.1091	0.6499	1.0150	234.1300	0.9767	1.4263
S3	✓	✓	0	–	4.1551	0.6590	1.0174	229.0889	0.9778	1.4269
S4	✓	✓	1	–	3.9369	0.6517	1.0212	230.8904	0.9762	1.4128
S5	✓	✓	2	–	3.9906	0.6455	1.0103	**227.9330**	0.9770	1.4238
S6	✓	✓	3	–	4.0560	0.6673	1.0257	232.1197	**0.9789**	1.4209
S7	✓	✓	0	✓	3.8901	0.6317	0.9975	230.6903	0.9731	1.4239
Ours	✓	✓	3	✓	**4.2938**	**0.6810**	**1.0401**	231.7368	0.9785	1.4059

best performance on both tasks. The higher mIoU indicates better preservation of semantic regions and scene structures, while the improved detection accuracy reflects retention of target-related and discriminative information.

4.4 Ablation Studies

The complete model consists of three modules: DHFE-PF, CSD-PR, and CGPO. DHFE-PF combines the DINOv3 hierarchical feature branch with the SPIB to integrate global visual priors and local details. CSD-PR performs progressive semantic diagnosis and modulation across fusion stages, while CGPO improves fusion quality through candidate generation and reward-driven policy optimization. Table 3 reports the ablation configurations. S1 and S2 use only SPIB and the DINOv3 hierarchical feature branch, respectively, while S3 combines both to form DHFE-PF. Based on S3, S4–S6 introduce one, two, and three stages of semantic modulation, with S6 using the full CSD-PR but without CGPO. S7 adds CGPO directly to DHFE-PF without CSD-PR. For configurations without CGPO, the model is trained by minimizing a conventional fusion loss. The complete model combines DHFE-PF, CSD-PR, and CGPO. Further ablation details are provided in Appendix A.9.

Table 2: Quantitative comparison of different fusion methods on semantic segmentation and object detection tasks.

Methods	mIoU ↑	$\mathrm{mAP}_{0.50:0.95}\uparrow$
MetaLFusion Li et al. (2021a)	0.5601	0.6312
DATFuse Tang et al. (2023)	0.5703	0.6363
TIMFusion Liu et al. (2024)	0.6143	0.6348
LUT-Fuse Yi et al. (2025)	0.6096	0.6387
SMLNet Kang et al. (2025)	0.5908	0.6455
FreeFusion Zhao et al. (2025)	0.6169	0.6501
DPOFusion Su et al. (2026)	0.5947	0.6180
DRFusion Shi et al. (2026)	0.6048	0.6400
Ours	**0.6293**	**0.6507**

Compared with S1, S2 performs better on Q_{MI}, $Q_{\mathrm{AB/F}}$, Q_{VIF}, and Q_{Y}. S3 further improves several metrics over S2 by combining global visual priors with local texture and edge information. S4–S6 progressively increase the number of semantic diagnosis stages. S6 achieves the best $Q_{\mathrm{AB/F}}$, Q_{VIF}, and Q_{Y} among these configurations, showing the benefit of multi-stage semantic refinement. Compared with S6, the complete model further improves Q_{MI}, $Q_{\mathrm{AB/F}}$, and Q_{VIF} by introducing CGPO. In contrast, applying CGPO directly to DHFE-PF in S7 degrades all six metrics compared with S3. This suggests that reward-driven candidate optimization is more effective on semantically refined fusion states, highlighting the complementarity between CSD-PR and CGPO.

5 Conclusion

This paper re-examines IR-VIS image fusion through semantic diagnosis and reward-driven optimization. Instead of relying only on fixed low-level visual constraints, we formulate fusion as a progressive process with intermediate-state diagnosis and candidate-based optimization. Frozen visual foundation models provide multi-scale visual priors, while vision-language embeddings diagnose and regulate intermediate fusion states. In addition, candidate generation, multidimensional rewards, and policy optimization enable learning from relative quality comparisons rather than predefined loss minimization. Experiments on multiple public datasets demonstrate consistent improvements in fusion quality and downstream semantic segmentation and object detection.

AI USE STATEMENT

In this work, we used large language models to polish the manuscript's language, check for linguistic errors, and assist in modifying portions of the code. The authors designed the overall code structure and wrote the main implementation. These language models were not used to generate experimental data. All AI-assisted content was reviewed by the authors, who take full responsibility for the entire work.

ETHICS STATEMENT

This work uses publicly available datasets for image fusion research. We comply with the terms of use of the datasets and pretrained models and appropriately cite their sources. Our research fully complies with the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The main paper describes the proposed method, datasets, and evaluation metrics. The appendix provides further implementation details, including module architectures, training configurations, and inference procedures, together with complete quantitative and qualitative results. We will release the code upon acceptance to support reproducibility.

REFERENCES

Haowen Bai, Jiangshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao Feng, and Shuang Xu. Task-driven image fusion with learnable fusion loss. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 7457–7468, 2025.

Hao Chen and Pramod K Varshney. A human perception inspired quality metric for image fusion based on regional information. *Information Fusion*, 8(2):193–207, 2007.

Jun Chen, Jianfeng Ding, and Jiayi Ma. Hitfusion: Infrared and visible image fusion for high-level vision tasks using transformer. *IEEE Transactions on Multimedia*, 26:10145–10159, 2024.

Yida Chen, Yafei Zhang, Huafeng Li, Zhengtao Yu, and Yu Liu. Joint multi-view embedding with progressive multi-scale alignment for unaligned infrared-visible image fusion. *Information Fusion*, 128:103960, 2026.

Chunyang Cheng, Tianyang Xu, Xiao-Jun Wu, Hui Li, Xi Li, Zhangyong Tang, and Josef Kittler. Textfusion: Unveiling the power of textual semantics for controllable image fusion. *Information Fusion*, 117:102790, 2025. ISSN 1566-2535.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8, 2023. URL `https://github.com/ultralytics/ultralytics`.

Huan Kang, Hui Li, Tianyang Xu, Xiao-Jun Wu, Rui Wang, Chunyang Cheng, and Josef Kittler. Smlnet: A spd manifold learning network for infrared and visible image fusion. *International Journal of Computer Vision*, 133(12):8642–8663, 2025.

Huafeng Li, Yueliang Cen, Yu Liu, Xun Chen, and Zhengtao Yu. Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion. *IEEE Transactions on Image Processing*, 30:4070–4083, 2021a.

Huafeng Li, Zengyi Yang, Yafei Zhang, Wei Jia, Zhengtao Yu, and Yu Liu. Mulfs-cap: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3673–3690, 2025.

Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019.

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, 2021b.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026a. URL https://arxiv.org/abs/2601.04720.

Xilai Li, Huichun Liu, Xiaosong Li, Tao Ye, Zhenyu Kuang, and Huafeng Li. Awm-fuse: Multi-modality image fusion for adverse weather via global and local text perception. *IEEE Transactions on Image Processing*, 35:5151–5164, 2026b.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33869–33880, June 2026c.

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)*, pp. 5802–5811, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8115–8124, October 2023.

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 Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2349–2369, 2025a.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2226–2235, 2025b.

Jinyuan Liu, Xingyuan Li, Qingyun Mei, Haoyuan Xu, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Bridging human evaluation to infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12322–12333, June 2026.

Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and meta-initialized deep model for image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6594–6609, 2024.

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

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.

Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics Letters*, 38(7):313–315, 2002.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

H.R. Sheikh and A.C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, 2006.

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

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel HAZIZA, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Weijian Su, Songqian Zhang, Yuqi Han, Jian Zhuang, Yongdong Huang, and Qiang Zhang. Fusion in your way: Aligning image fusion with heterogeneous demands via direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41499–41509, June 2026.

Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83-84:79–92, 2022.

Linfeng Tang, Yeda Wang, Zhanchuan Cai, Junjun Jiang, and Jiayi Ma. Controlfusion: A controllable image fusion network with language-vision degradation prompts. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 118304–118327, 2025.

Wei Tang, Fazhi He, Yu Liu, Yansong Duan, and Tongzhen Si. Datfuse: Infrared and visible image fusion via dual attention transformer. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3159–3172, 2023.

Alexander Toet. TNO image fusion dataset, 2014. URL `https://doi.org/10.6084/m9.figshare.1008029.v1`. Version 1.

Yuhao Wang, Lingjuan Miao, Zhiqiang Zhou, Lei Zhang, and Yajun Qiao. Infrared and visible image fusion with language-driven loss in clip embedding space. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, pp. 1443—-1451, 2025a.

Zeyu Wang, Libo Zhao, Jizheng Zhang, Rui Song, Haiyu Song, Jiana Meng, and Shidong Wang. Multi-text guidance is important: Multi-modality image fusion via large generative vision-language model. *International Journal of Computer Vision*, 133(7):4646–4668, 2025b.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.

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)*, volume 34, pp. 12077–12090, 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.

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

Cui Yang, Jian-Qi Zhang, Xiao-Rui Wang, and Xin Liu. A novel similarity based quality metric for image fusion. *Information Fusion*, 9(2):156–160, 2008.

Zengyi Yang, Yu Liu, Juan Cheng, Zhiqin Zhu, Yafei Zhang, and Huafeng Li. Customized fusion: A closed-loop dynamic network for adaptive multi-task-aware infrared-visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 188–198, June 2026.

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 (CVPR)*, pp. 27026–27035, 2024.

Xunpeng Yi, Yibing Zhang, Xinyu Xiang, Qinglong Yan, Han Xu, and Jiayi Ma. Lut-fuse: Towards extremely fast infrared and visible image fusion via distillation to learnable look-up tables. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14559–14568, 2025.

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

Wenda Zhao, Hengshuai Cui, Haipeng Wang, You He, and Huchuan Lu. Freefusion: Infrared and visible image fusion via cross reconstruction learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):8040–8056, 2025.

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, 2023.

A Appendix

This appendix provides additional methodological details, analyses, and experimental results that complement the main paper. It includes implementation details, further method analysis, additional experiments, ablation studies, and limitation and future-work discussions.

A.1 Detailed Architecture of DHFE-PF

A.1.1 Input preprocessing for DINOv3

Single-channel inputs are replicated across three channels and preprocessed using the image processor associated with the pretrained DINOv3 checkpoint, with resizing disabled in our configuration. The IR and VIS inputs are then fed separately into the shared frozen DINOv3-ConvNeXt encoder for hierarchical feature extraction.

A.1.2 Spatial Prior Injection Branch

The Spatial Prior Injection Branch (SPIB) is used to compensate for local textures, edges, and high-frequency information that may be weakened by progressive downsampling in the frozen DINOv3-ConvNeXt encoder. Specifically, we use the DINOv3-ConvNeXt-Base-Pretrain-Lvd1689m checkpoint as the frozen encoder.

Given the registered IR and VIS images, SPIB first concatenates them along the channel dimension and feeds them into a shallow mapping network $\mathcal{C}_{\text{prior}}$. The network consists of two 3×3 convolutional layers with GroupNorm and LeakyReLU. The resulting feature serves as the base spatial prior.

To align the spatial prior with the hierarchical features extracted by DINOv3-ConvNeXt, the base feature is projected to four spatial scales through scale-specific mappings ψ_k, $k \in \{0, 1, 2, 3\}$. Each ψ_k consists of a 3×3 convolution followed by a 1×1 convolution, together with GroupNorm and LeakyReLU. Bilinear interpolation is used to match the target spatial resolution. The resulting features $\{\mathbf{S}_0, \mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3\}$ provide local spatial priors for the subsequent progressive fusion process.

A.1.3 Multi-source Collaborative Fusion Module

Within the Multi-source Collaborative Fusion Module (MSCM), the Cross-modal Fusion Block (CFB) at each scale consists of two 3×3 convolutional layers with GroupNorm and LeakyReLU. It first integrates the IR and VIS representations at the same hierarchical level.

For cross-scale interaction, the higher-level fused feature is bilinearly interpolated to the spatial size of the current adaptation feature $\mathbf{S}_k$. A Channel Alignment Projection (CAP) $\mathcal{P}_k$, implemented using a 1×1 convolution, then aligns its channel dimension before feature addition. At the highest scale ($k = 3$), no higher-level fused feature is available, and this cross-scale injection is omitted.

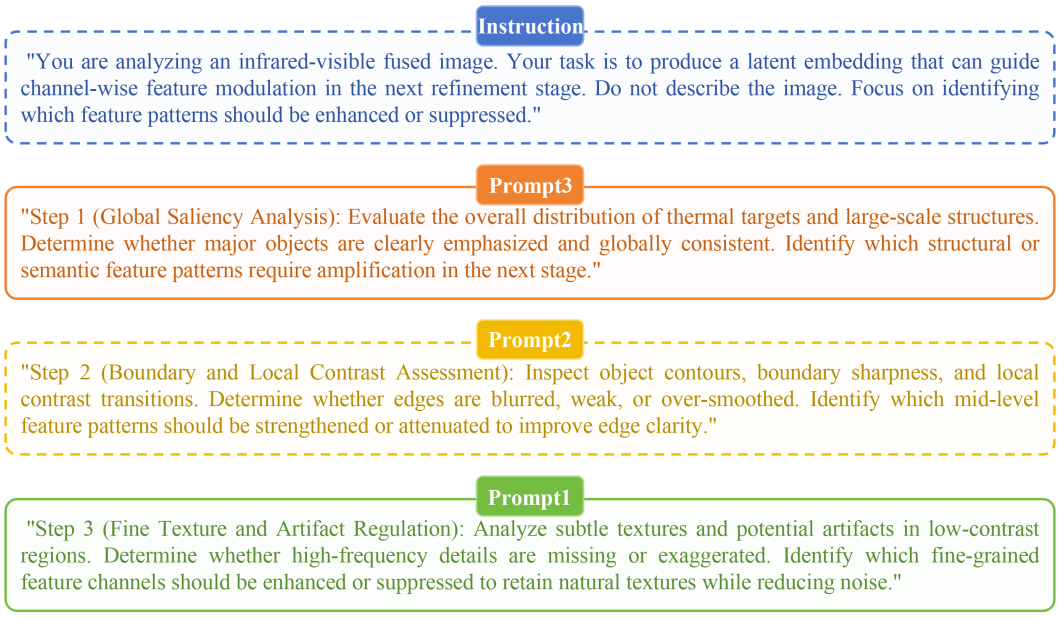


Figure 5: The shared instruction and stage-specific prompts used for semantic diagnosis in CSD-PR.

After cross-modal fusion, cross-scale injection, and local prior compensation, the aggregated representation is refined by a Feature Refinement Projection (FRP). Each FRP consists of a 3×3 convolution followed by GroupNorm and LeakyReLU.

A.1.4 TPHF Implementation

The Top-down Progressive Hierarchical Fusion (TPHF) pathway further aggregates the refined multi-scale features produced by MSCM. Different from the preceding cross-scale injection, which operates at the native resolution of each hierarchical level, TPHF first maps the refined features to the input spatial resolution and then performs top-down residual aggregation.

The Channel Compression and Refinement (CCR) module $\mathcal{R}_k$ used at each scale consists of a 3×3 convolution, GroupNorm, and LeakyReLU. Bilinear interpolation is used to resize $\mathbf{X}_k^{raw}$ to the input resolution before residual aggregation. At the input scale, the full-resolution spatial prior $\mathbf{S}_0$ is additionally introduced to compensate for fine-grained details.

A.2 Additional Details of CSD-PR

A.2.1 Stage-specific Semantic Diagnosis

Semantic diagnosis is performed at Stages 3, 2, and 1. Stage 0 does not invoke Qwen3-VL-Embedding again and only receives the semantic modulation propagated from Stage 1.

The same fixed instruction is shared across all diagnostic stages, while the stage-specific prompt changes according to the semantic characteristics of each hierarchical level. The Stage 3 prompt focuses on salient targets and global structural consistency. The Stage 2 prompt emphasizes object boundaries, regional contrast, and structural continuity. The Stage 1 prompt focuses on textural details, local edges, and artifact suppression. The fixed instruction and stage-specific prompts are shown in Figure 5.

This hierarchical prompt design allows the semantic diagnostic criterion to adapt to the information represented at different fusion stages rather than applying an identical prompt at all levels.

Fixed prompts guide Qwen3-VL-Embedding toward stage-specific diagnostic priorities, while visual alignment and learned projections adapt its embeddings to image content, aiming to reduce sensitivity to exact prompt wording.

A.2.2 GLOBAL VISUAL REPRESENTATION

Let $\mathbf{h}_g^m = \{\mathbf{h}_g^m[\tau]\}_{\tau=1}^T$ denote the global feature sequence of modality m, where T denotes the number of elements in the sequence. The joint global visual representation used for visual-semantic alignment is computed as

$$\mathbf{v}_g = \frac{1}{2T} \sum_{\tau=1}^{T} \left(\mathbf{h}_g^{ir}[\tau] + \mathbf{h}_g^{vis}[\tau] \right). \tag{13}$$

This representation combines the global structural information of both source modalities and serves as a visual reference for aligning the semantic representation generated by Qwen3-VL-Embedding.

A.2.3 VISUAL-SEMANTIC ALIGNMENT DETAILS

We use the 2B variant of Qwen3-VL-Embedding, which outputs a 2048-dimensional semantic representation $\mathbf{l}_k$ at each diagnostic stage. The projection $\mathbf{W}_q$ first maps $\mathbf{l}_k$ into the visual embedding space. The stage-specific projection $\mathbf{W}_{proj}^{(k)}$ then maps the aligned representation to the channel dimension of the next finer-scale feature $\mathbf{Y}_{k-1}$. Qwen3-VL-Embedding provides embeddings rather than textual critiques or explicit feature corrections; our projection heads align these embeddings with visual features to guide channel modulation.

The centered channel gate $\boldsymbol{\gamma}_k = \sigma(\mathbf{f}_k) - 0.5$ lies in $[-0.5, 0.5]$. Compared with a conventional Sigmoid gate, the centered formulation allows both positive and negative modulation. It can therefore enhance informative channel responses while suppressing channels associated with semantic degradation.

During Progressive Semantic Channel Modulation (PSCM), $\boldsymbol{\gamma}_k$ is broadcast along the spatial dimensions to match $\mathbf{Y}_{k-1}$. The modulated feature is then used as the current diagnostic state at the next finer stage.

A.2.4 SHARED DECODER ARCHITECTURE

The lightweight decoder $g(\cdot)$ is shared between intermediate-state reconstruction in CSD-PR and candidate image reconstruction in CGPO. It consists of two 3×3 convolutional layers with GroupNorm and LeakyReLU, followed by a 1×1 convolution that produces a single-channel fused image.

Sharing the decoder keeps intermediate semantic diagnosis and candidate evaluation in the same image reconstruction space and avoids introducing separate decoders for the two processes.

Gradient flow through the shared decoder. Intermediate diagnostic images are detached and converted to 8-bit PIL images before being processed by Qwen3-VL-Embedding under `torch.no_grad()`. Therefore, this diagnostic pathway does not backpropagate to the shared decoder. The decoder instead receives gradients through the differentiable rewards of the final candidate images. Its updated parameters are reused for intermediate reconstruction in subsequent forward passes.

A.3 ADDITIONAL DETAILS OF CGPO

A.3.1 POLICY HEAD ARCHITECTURE

The candidate policy head π_θ consists of two 3×3 convolutional layers. The first layer is followed by GroupNorm and LeakyReLU, while the second predicts a residual feature offset. This offset is added to the semantically refined feature $\mathbf{Z}$ to form the mean of the candidate distribution.

The standard deviation of the isotropic Gaussian policy is fixed to $\sigma = 0.1$ in all experiments. The fixed variance restricts candidate exploration to a local neighborhood of the current fusion state and avoids introducing additional variance-prediction parameters.

Gaussian perturbation is applied only during training. During inference, no random perturbation is introduced, and only the reference candidate corresponding to the policy mean is decoded.

A.3.2 CANDIDATE LOG-PROBABILITY

Assuming independent Gaussian perturbations over channel and spatial dimensions, the normalized log-probability of each candidate is computed as

$$\log \pi_\theta\big(\mathbf{a}^{(n)} \mid \mathbf{Z}\big) = \frac{1}{CHW} \sum_{c,h,w} \log \mathcal{N}\left(a_{c,h,w}^{(n)}; \mu_{c,h,w}, \sigma^2\right), \quad n = 0, \ldots, N-1. \tag{14}$$

Here, $\mu_{c,h,w}$ denotes the policy mean at location (c, h, w). The normalization by CHW reduces the dependence of the log-probability scale on feature dimensionality. In the candidate objective, $\log \pi_\theta$ is shorthand for the joint Gaussian log density divided by CHW.

Differentiation of the candidate optimization loss. CGPO jointly evaluates the unperturbed reference candidate and its perturbed variants under the same reward criteria. All candidates contribute to the group-mean baseline and the candidate optimization loss $\mathcal{L}_{\mathrm{RL}}$ defined in Eq. 12. This loss weights their group-relative advantages by the normalized candidate log densities, aiming to improve the reference candidate's reward relative to its perturbed alternatives and guide the optimization of fusion quality. Candidate features, rewards, and the group-mean baseline remain in the computation graph. Backpropagation accounts for each reward's direct contribution to $\mathcal{L}_{\mathrm{RL}}$ and its indirect contribution through the group-mean baseline. The resulting reward gradients pass through the reconstructed candidate images and the shared decoder to the candidate features, then through the policy mean to the candidate-generation policy and connected upstream fusion modules. During training, $\mathcal{L}_{\mathrm{RL}}$ is averaged over input samples in each batch.

A.3.3 GROUP-MEAN BASELINE AND ADVANTAGE ESTIMATION

For each input sample, all N candidates, including the reference candidate, participate in reward comparison and group-mean baseline estimation. The relative advantage of each candidate is computed by subtracting the group-mean reward from its reward. The group mean serves as a comparison baseline for the constructed candidate group, rather than an unbiased estimate of the expected reward under the Gaussian policy.

Gradient of the candidate objective. With fixed variance and reparameterized candidates, the total derivative of each candidate log density with respect to the policy mean is zero. Gradients of $\mathcal{L}_{\mathrm{RL}}$ flow through candidate rewards and the differentiable group-mean baseline, which centers the log-density coefficients in the reward gradient. The resulting update optimizes a differentiable within-group reward-comparison objective. The GRPO inspiration concerns within-group reward comparison; our update differentiates through candidate rewards rather than using a REINFORCE score-function estimator.

A.3.4 REWARD COMPUTATION DETAILS

Since the fused image is single-channel, the VIS image is converted to its luminance component before reward computation. We denote the luminance image as $\mathbf{I}_{vis}^{Y}$. For pixel response consistency, the reference image is constructed using the stronger response of the two source modalities:

$$\mathbf{I}_{ref} = \max(\mathbf{I}_{ir}, \mathbf{I}_{vis}^{Y}). \tag{15}$$

For structural similarity, SSIM is computed between each candidate fused image and the IR image, as well as between the candidate and the VIS luminance image. The resulting structural reward is restricted to $[\varepsilon, 1]$. The lower bound ε prevents the multiplicative composite reward from approaching zero when one reward component becomes very small, and $\varepsilon = 10^{-4}$ in our experiments.

For gradient preservation, Sobel operators are used to extract gradient magnitudes from the candidate image, the IR image, and the VIS luminance image. The reference gradient is constructed using the pixel-wise maximum of the two source gradients:

$$g_{ref} = \max(g_{ir}, g_{vis}). \tag{16}$$

This design encourages the fused result to preserve salient edge responses from both modalities.

Table 4: Quantitative comparison of different fusion methods on the MSRS, M3FD, and FMB datasets. The best and second-best results within each dataset are highlighted in **pink** and blue, respectively.

Methods	$Q_{\mathrm{MI}}\uparrow$	$Q_{\mathrm{AB/F}}\uparrow$	$Q_{\mathrm{VIF}}\uparrow$	$Q_{\mathrm{CV}}\downarrow$	$Q_{\mathrm{Y}}\uparrow$	$Q_{\mathrm{SSIM}}\uparrow$
MSRS						
MetaLFusion Li et al. (2021a)	2.4349	0.2618	0.4232	658.1897	0.7615	1.2942
DATFuse Tang et al. (2023)	3.7634	0.6309	0.8808	322.1601	0.9395	1.2345
TIMFusion Liu et al. (2024)	3.5217	0.5032	0.8177	424.2178	0.9070	1.2819
LUT-Fuse Yi et al. (2025)	3.6258	0.6079	0.9743	287.1405	0.9548	1.3860
SMLNet Kang et al. (2025)	3.3045	0.5418	0.8528	447.7691	0.9384	1.3705
FreeFusion Zhao et al. (2025)	1.4098	0.3972	0.5085	900.7477	0.6849	0.8387
DPOFusion Su et al. (2026)	1.6436	0.2104	0.7297	905.6706	0.5681	0.8197
DRFusion Shi et al. (2026)	3.0273	0.4654	0.7921	255.9119	0.9583	1.3597
Ours	**4.2938**	**0.6810**	**1.0401**	**231.7368**	**0.9785**	**1.4059**
M3FD						
MetaLFusion Li et al. (2021a)	3.7181	0.4315	0.7487	727.8273	0.8346	1.3843
DATFuse Tang et al. (2023)	3.9986	0.4884	0.7364	593.4340	0.8388	1.3794
TIMFusion Liu et al. (2024)	3.9427	0.5359	0.7481	868.0657	0.8800	1.3689
LUT-Fuse Yi et al. (2025)	3.9837	0.4863	0.8015	642.2529	0.8393	1.3958
SMLNet Kang et al. (2025)	2.8667	0.4967	0.7630	603.2871	0.8520	1.3817
FreeFusion Zhao et al. (2025)	2.3569	0.5071	0.7373	760.5181	0.8409	1.3014
DPOFusion Su et al. (2026)	1.8361	0.1952	0.6613	954.1886	0.6286	1.2261
DRFusion Shi et al. (2026)	3.6875	0.4073	0.7717	626.2359	0.8612	1.3797
Ours	**4.1351**	**0.6101**	**0.8923**	**572.1672**	**0.9080**	**1.4061**
FMB						
MetaLFusion Li et al. (2021a)	3.9883	0.4845	0.8138	321.2907	0.8701	1.4407
DATFuse Tang et al. (2023)	4.0752	0.5695	0.8180	257.8650	0.8741	1.4303
TIMFusion Liu et al. (2024)	3.7427	0.5813	0.7244	249.8619	0.9085	1.4228
LUT-Fuse Yi et al. (2025)	3.8829	0.5251	0.8635	269.9742	0.8778	1.4536
SMLNet Kang et al. (2025)	2.8909	0.6104	0.8267	229.1527	0.9121	1.4400
FreeFusion Zhao et al. (2025)	2.4963	0.5414	0.7579	525.7139	0.8713	1.3692
DPOFusion Su et al. (2026)	1.8428	0.2167	0.6838	569.8307	0.6167	1.2975
DRFusion Shi et al. (2026)	3.5768	0.4515	0.8337	262.4585	0.8931	1.4421
Ours	**4.2359**	**0.6672**	**0.9729**	**205.5012**	**0.9541**	**1.4609**

A.4 Implementation Details

The proposed method is implemented in Python 3.10.6 with PyTorch 2.6.0 and trained on a single NVIDIA RTX 4090 GPU. We train the fusion model for 1,000 epochs with a batch size of 4. Input images are normalized to $[0, 1]$ and randomly cropped into 256×256 patches during training. We use the Adam optimizer with an initial learning rate of 1×10^{-4}, which is gradually decayed to 1×10^{-6} using cosine annealing. For candidate generation, we produce four fusion candidates for each input sample, including one reference candidate and three perturbed candidates. With a batch size of 4, each forward pass therefore yields 16 candidate results. Reward evaluation and advantage estimation are performed independently within the four candidates of each sample, and the resulting advantages are used for policy optimization. Notably, all models are deployed locally. During inference, random perturbation is disabled, and $\mathbf{I}_f^{(0)}$ is used as the final fused output.

A.5 Training and Inference Details

During training, DINOv3-ConvNeXt and Qwen3-VL-Embedding remain frozen, while SPIB, the lightweight fusion modules, visual-semantic alignment and modulation layers, shared decoder, and policy head are optimized. CSD-PR progressively diagnoses and regulates intermediate fusion

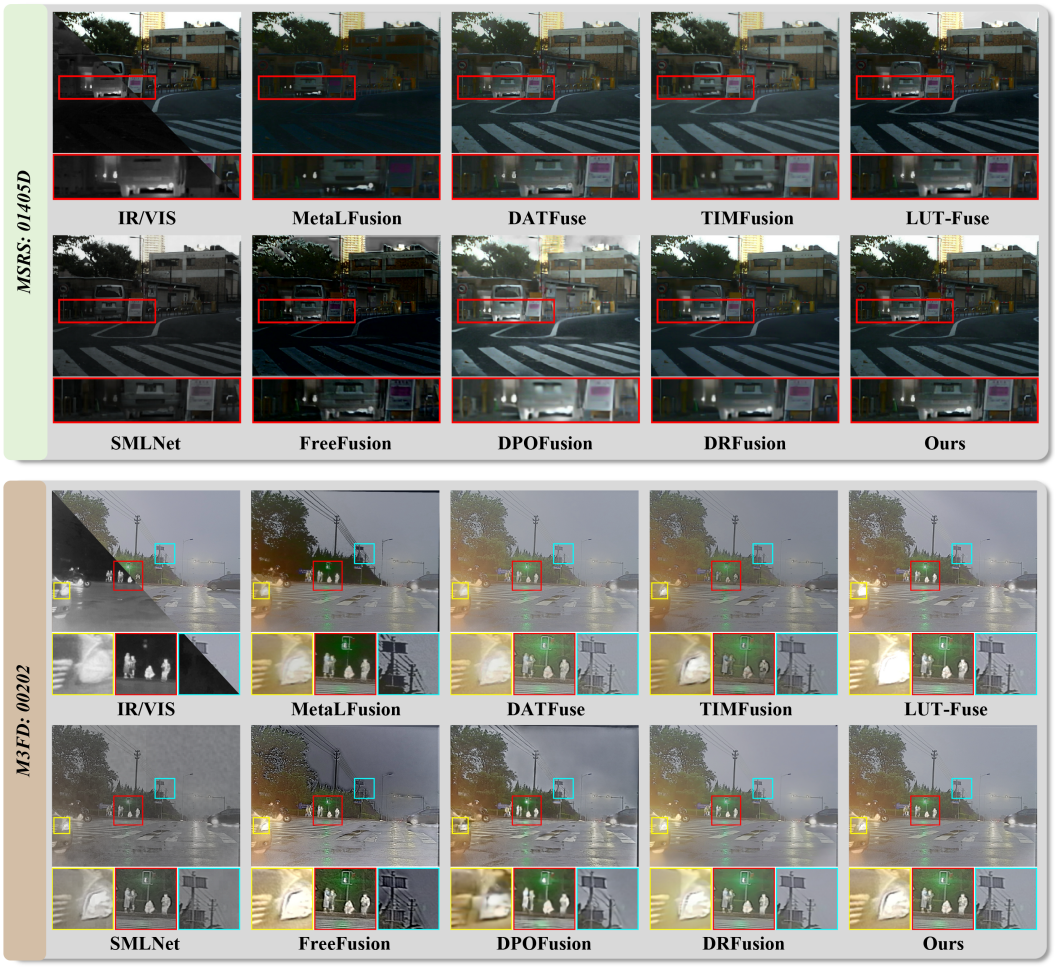


Figure 6: Qualitative comparison of different fusion methods on the MSRS and M3FD datasets.

states. CGPO then generates multiple candidate features through Gaussian perturbation and evaluates their relative quality using the multidimensional reward.

During inference, Gaussian perturbation, candidate comparison, reward computation, and policy optimization are removed. CSD-PR remains active to produce the semantically refined feature $\mathbf{Z}$. The policy mean is used as the reference candidate and decoded by the shared decoder to obtain the final fused image.

A.6 Detailed Quantitative Results

Table 4 reports the complete quantitative results of our method and competing methods on MSRS, M3FD, and FMB, complementing the average results presented in Table 1 of the main paper. The reported mean values are truncated rather than rounded. For all datasets, we use the official training and test splits without random repartitioning to ensure fairness and reproducibility. Our method achieves the best performance on all six metrics across all three datasets, demonstrating its overall effectiveness in information preservation, edge detail recovery, and structural consistency. Notably, although trained only on the MSRS training set, our method maintains its leading performance on M3FD and FMB. These results support its ability to generalize across datasets with different scenes and data distributions, further demonstrating its effectiveness in producing high-quality fused images.

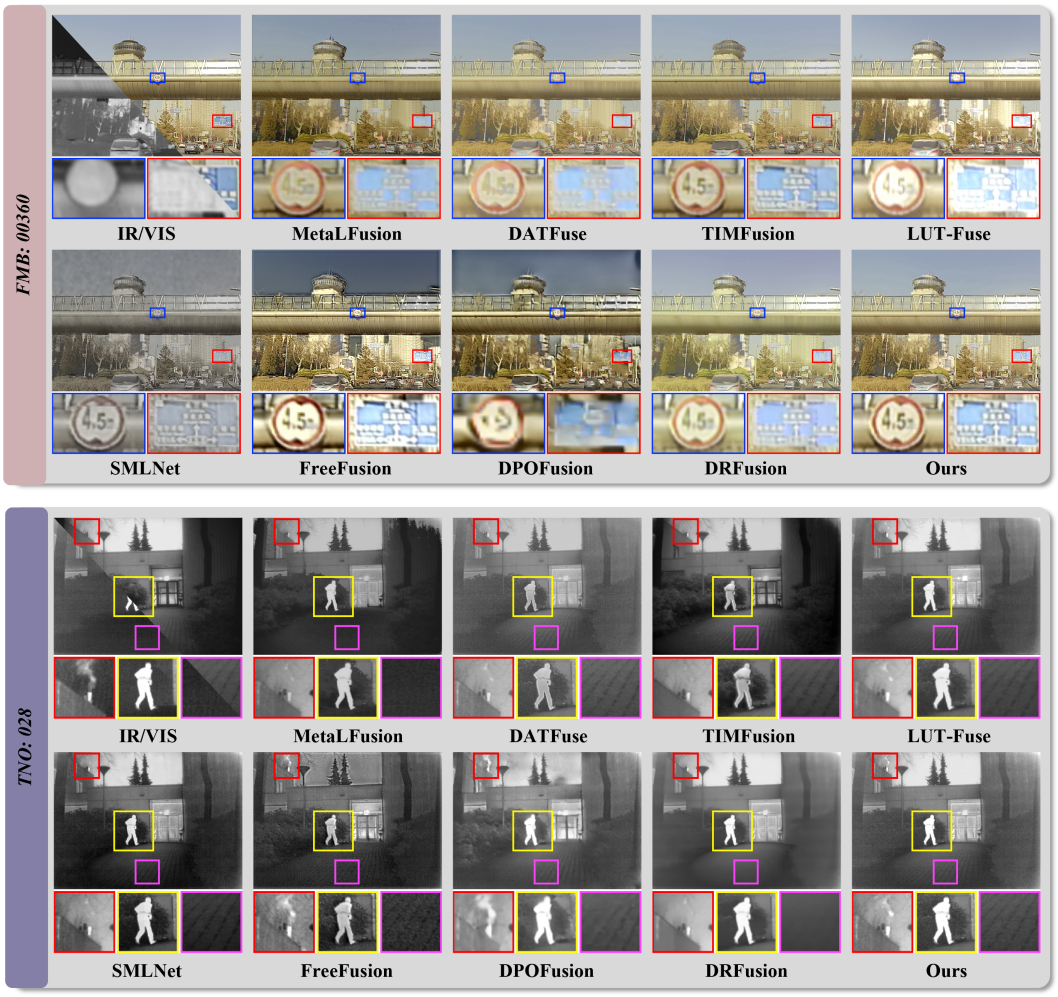


Figure 7: Qualitative comparison of different fusion methods on the FMB and TNO datasets.

A.7 Detailed Qualitative Results

We provide additional qualitative comparisons on MSRS, M3FD, and FMB to complement the results in the main paper. We also include TNO (Toet, 2014) as an additional dataset to further examine the performance of our method across different scenes.

MSRS: Figure 6 presents qualitative comparisons on MSRS. Our method preserves salient infrared targets while retaining visible texture and structural details. In the highlighted region, the truck remains prominent, with clearer details on its front and the adjacent sign. While some competing methods produce dim targets or lose local details when increasing brightness, our method enhances target visibility while preserving fine structures, achieving a favorable balance between thermal saliency and detail preservation.

M3FD: Figure 6 presents qualitative comparisons on M3FD. Our method produces visually coherent results with salient pedestrians and well-preserved infrared information, while retaining fine wire structures in the enlarged region. LUT-Fuse also achieves competitive results, but its strong brightness enhancement obscures local details in the headlight region. Overall, our method better balances target saliency and detail preservation, maintaining clearer structures in bright regions.

FMB: Figure 7 presents qualitative comparisons on FMB. Our method maintains natural contrast in the daytime scene, avoiding excessive brightness enhancement and the associated loss of detail.

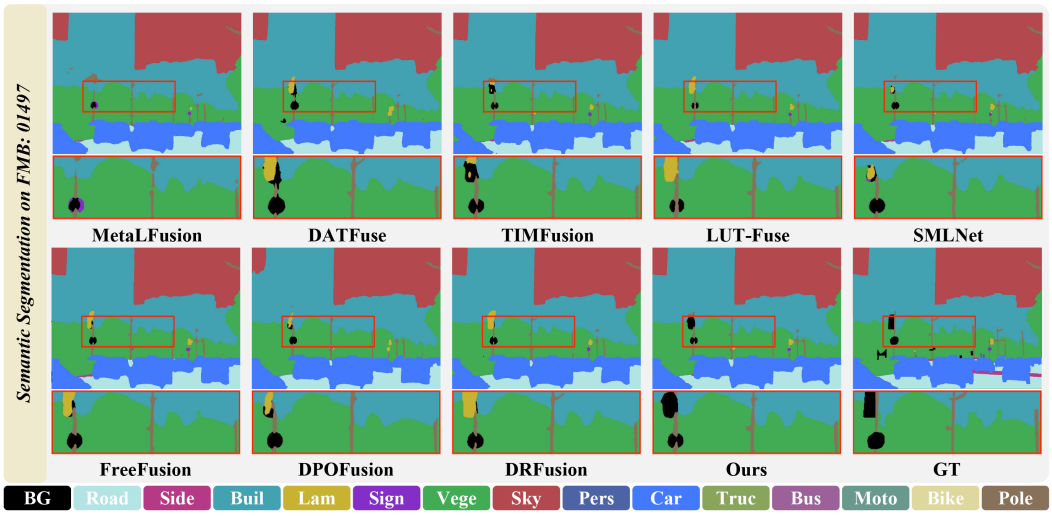


Figure 8: Qualitative comparison of semantic segmentation results on the FMB dataset.

Table 5: Semantic segmentation results on the FMB dataset using SegFormer-B5. Per-class IoU and overall mIoU are reported in percentages. The best and second-best results are highlighted in **pink** and blue, respectively.

Methods	BG	Road	Side	Build	Lamp	Sign	Vege	Sky	Pers	Car	Truck	Bus	Moto	Pole	mIoU
MetaLFusion Li et al. (2021a)	15.27	86.64	36.54	81.57	21.46	62.38	84.54	93.94	64.41	81.97	26.55	56.12	29.08	43.75	56.01
DATFuse Tang et al. (2023)	16.11	87.63	39.46	81.85	24.61	68.45	85.37	93.34	65.71	82.84	16.72	55.51	36.84	43.92	57.03
TIMFusion Liu et al. (2024)	15.80	88.47	45.92	81.73	34.35	73.02	85.64	93.85	64.33	82.97	**45.61**	59.70	42.39	46.19	61.43
LUT-Fuse Yi et al. (2025)	15.69	87.69	43.02	83.09	33.74	71.98	85.93	93.58	67.60	82.44	31.19	**72.82**	37.81	46.93	60.96
SMLNet Kang et al. (2025)	16.28	88.42	45.89	81.83	37.00	71.71	85.83	93.34	67.22	80.13	30.74	40.78	41.14	46.72	59.08
FreeFusion Zhao et al. (2025)	**16.69**	88.31	44.54	**84.09**	34.54	**74.12**	86.96	94.38	68.40	82.85	33.11	63.45	45.14	47.09	61.69
DPOFusion Su et al. (2026)	15.29	87.59	46.02	81.51	**40.31**	71.14	85.50	93.63	63.99	79.88	28.79	65.98	30.10	42.85	59.47
DRFusion Shi et al. (2026)	16.12	88.13	45.06	82.62	36.12	73.74	86.25	93.78	66.43	**83.05**	29.79	64.03	35.21	46.44	60.48
Ours	16.48	**88.59**	**46.61**	83.97	37.77	73.77	**87.04**	**94.46**	**68.48**	83.02	35.29	70.16	**47.47**	**47.89**	**62.93**

In the enlarged regions, the distant traffic signs retain clearer text and finer structures. Overall, our method preserves source image details while maintaining a natural visual appearance.

TNO: We provide additional qualitative results on TNO to further examine the effectiveness of our method in challenging scenes. As shown in Figure 7, our method preserves the salient infrared pedestrian in the center while retaining subtle smoke details in the upper-left region. Meanwhile, the pavement textures remain visible, supporting balanced overall visual quality. These observations further demonstrate the effectiveness of our method in preserving both thermal target saliency and fine background details.

A.8 Detailed Downstream Task Results

A.8.1 Semantic Segmentation

Table 5 reports the per-class semantic segmentation results obtained using SegFormer-B5 on the official FMB test set, which contains 280 infrared–visible image pairs. Class definitions follow the official FMB dataset, but the Bicycle class is absent from the official test set. All methods follow the official training and test splits. For each fusion method, we generate fused images for both splits, train a separate SegFormer-B5 model on its fused training images using the same architecture and training hyperparameters, and evaluate the trained model on its corresponding fused test images. All SegFormer-B5 models are trained for 30 epochs. Our method achieves the highest mIoU of 62.93%. It obtains the best IoU in seven of the fourteen categories and the second-best IoU in the

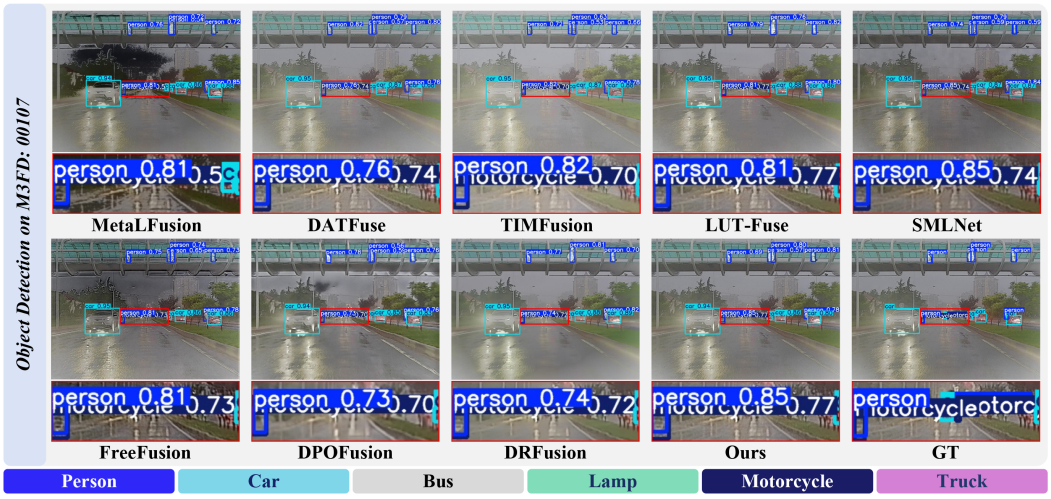


Figure 9: Qualitative comparison of object detection results on the M3FD dataset.

Table 6: Object detection results on the M3FD dataset using YOLOv8s. Per-class results are measured by $\mathrm{AP}_{50:95}$, and All denotes the overall $\mathrm{mAP}_{50:95}$. Higher values indicate better performance. The best and second-best results are highlighted in **pink** and blue, respectively. Tied results receive the same highlighting.

Methods	Person	Car	Bus	Lamp	Motorcycle	Truck	All
MetaLFusion Li et al. (2021a)	0.578	0.740	0.787	0.490	0.552	0.641	0.6312
DATFuse Tang et al. (2023)	0.586	0.737	0.814	0.534	0.504	0.643	0.6363
TIMFusion Liu et al. (2024)	0.559	0.752	0.819	0.538	0.513	0.626	0.6348
LUT-Fuse Yi et al. (2025)	0.589	0.744	0.821	0.526	0.518	0.635	0.6387
SMLNet Kang et al. (2025)	0.583	0.746	0.823	0.522	**0.563**	0.635	0.6455
FreeFusion Zhao et al. (2025)	0.585	**0.754**	0.818	**0.552**	0.527	**0.665**	0.6501
DPOFusion Su et al. (2026)	0.549	0.730	0.793	0.529	0.473	0.634	0.6180
DRFusion Shi et al. (2026)	0.587	0.744	0.816	0.528	0.541	0.624	0.6400
Ours	**0.593**	**0.754**	**0.829**	0.538	0.531	0.659	**0.6507**

remaining seven. These results suggest that our fused images preserve semantic information useful for segmentation and support consistent performance across different categories.

Figure 8 presents qualitative semantic segmentation results on FMB. In the highlighted region, our result more closely matches the ground-truth labels, reducing the misclassification of background regions as lamps while preserving thin pole structures. These results suggest that our semantic diagnosis and reward-driven optimization help retain structural and semantic cues for downstream segmentation.

A.8.2 Object Detection

Table 6 reports per-class YOLOv8s detection results on the 300 pairs of infrared and visible images officially provided for M3FD fusion evaluation. All methods use the same training and test splits. For each fusion method, we generate fused images for both splits, train a separate YOLOv8s model on its fused training images using the same architecture and training hyperparameters, and evaluate the trained model on its corresponding fused test images. All YOLOv8s models are trained for 300 epochs. Training separate detectors may narrow the gap through adaptation to each method's fused images. It is worth noting that all experiments are conducted using the same hardware, model configuration, and random seed to ensure a fair comparison. Under this consistent evaluation setting, our method achieves the highest overall $\mathrm{mAP}_{50:95}$ of 0.6507, slightly exceeding FreeFusion (0.6501). It obtains the best results for Person and Bus, ties for the best result for Car, and achieves

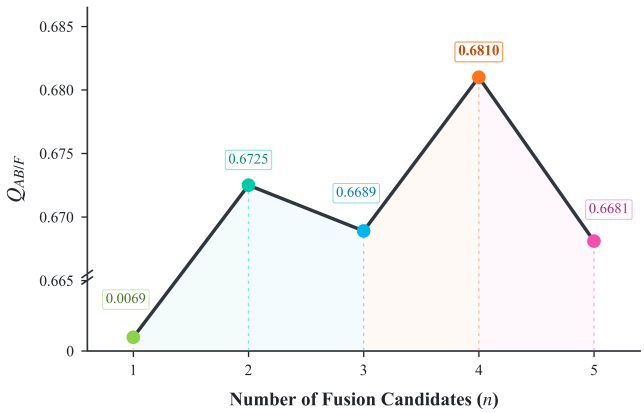


Figure 10: Effect of the number of candidate fusion images on performance.

the second-best results for Lamp and Truck. Although it does not lead in every category, these results suggest that our fused images effectively preserve object-related information useful for downstream detection.

Figure 9 presents qualitative object detection results on M3FD. Our fused images support the detection of small, distant objects, including pedestrians and vehicles. In the highlighted region, our method achieves a pedestrian detection confidence of 0.85, exceeding most compared methods. These results suggest that our semantic diagnosis and reward-driven optimization help preserve object-related cues for detection in challenging scenes.

A.9 Ablation Studies

A.9.1 Experimental Settings

For ablation configurations without CGPO (S1–S6 in Table 3), candidate generation and policy optimization are omitted, while all other operations of each configuration remain unchanged. The final fusion feature $\mathbf{Z}$ is directly fed into the decoder $g(\cdot)$ to produce a single fused image. The model is trained by minimizing a conventional fusion loss computed on this output, $\mathcal{L}_{\text{fusion}} = \mathcal{L}_{\text{pix}} + \mathcal{L}_{\text{SSIM}} + \mathcal{L}_{\text{grad}}$, where $\mathcal{L}_{\text{pix}}$, $\mathcal{L}_{\text{SSIM}}$, and $\mathcal{L}_{\text{grad}}$ denote the pixel consistency loss, SSIM-based structural loss, and gradient preservation loss, respectively, with equal weights of $1:1:1$.

Configurations with CGPO are trained solely using $\mathcal{L}_{\text{RL}}$, without additional fusion, reconstruction, or other auxiliary losses, whereas configurations without CGPO use only $\mathcal{L}_{\text{fusion}}$.

A.9.2 Interpretation of Module Ablations

Since disabling CGPO also replaces $\mathcal{L}_{\text{RL}}$ with $\mathcal{L}_{\text{fusion}}$, these comparisons evaluate candidate generation and its associated optimization objective jointly. Under this protocol, S7 underperforms S3 on all six metrics, whereas the complete model improves Q_{MI}, $Q_{\text{AB/F}}$, Q_{VIF}, and Q_{CV} over S6, with decreases in Q_{Y} and Q_{SSIM}. Increasing the number of diagnostic stages likewise does not yield monotonic improvements across all metrics. These observations suggest that the gains depend on the interaction between CSD-PR and CGPO and involve trade-offs among quality criteria. The complete model achieves the highest Q_{MI}, $Q_{\text{AB/F}}$, and Q_{VIF} in Table 3, supporting the combined design for information and edge preservation.

A.10 Model Configuration and Design Analysis

A.10.1 Analysis of the Number of Candidate Fused Images

The number of candidate fused images, N, controls the number of candidates explored around the current fusion state. It also affects GPU memory usage and training cost. Due to the 24 GB mem-

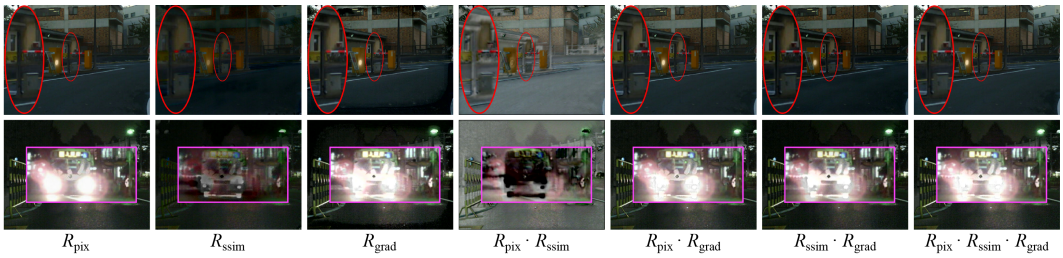


Figure 11: Qualitative comparison of different reward combinations on the MSRS dataset.

Table 7: Quantitative comparison of different reward combinations on the MSRS dataset. Overall denotes the average normalized ranking score across six metrics, where the best rank is assigned 1 and the worst rank is assigned 0. A higher Overall score indicates better comprehensive performance. The best and second-best results are highlighted in **pink** and **blue**, respectively.

Reward Setting	$Q_{\text{MI}} \uparrow$	$Q_{\text{AB/F}} \uparrow$	$Q_{\text{VIF}} \uparrow$	$Q_{\text{CV}} \downarrow$	$Q_{\text{Y}} \uparrow$	$Q_{\text{SSIM}} \uparrow$	Overall↑
R_{pix}	**5.1042**	0.6690	1.0020	249.25	0.9753	1.3921	0.639
R_{ssim}	1.6659	0.2112	0.5951	3486.84	0.8431	1.4039	0.194
R_{grad}	2.4249	0.5974	0.9266	329.10	0.9488	1.2335	0.306
$R_{\text{pix}} \cdot R_{\text{ssim}}$	1.3384	0.2155	0.2844	2862.24	0.5570	0.8062	0.056
$R_{\text{pix}} \cdot R_{\text{grad}}$	4.7354	0.6807	1.0344	234.91	0.9775	1.3909	0.722
$R_{\text{ssim}} \cdot R_{\text{grad}}$	3.0543	0.6698	0.9769	**212.81**	0.9674	**1.4272**	0.694
Ours $(R_{\text{pix}} \cdot R_{\text{ssim}} \cdot R_{\text{grad}})$	4.2938	**0.6810**	**1.0401**	231.74	**0.9785**	1.4059	**0.889**

ory limit of a single NVIDIA RTX 4090 GPU, we set the maximum candidate number to $N = 5$. Figure 10 shows the model performance under different candidate numbers. While optimizing with CGPO under the $N = 1$ setting, relative quality comparison among candidates cannot be established, and both the relative advantage and $\mathcal{L}_{\text{RL}}$ are identically zero, providing no training gradient from this objective. Therefore, the model fails to produce meaningful fused images. When $N > 1$, the model can estimate relative advantages from multiple candidates and optimize the generation strategy based on their quality differences. This generally leads to better edge preservation, as reflected by higher $Q_{\text{AB/F}}$ scores. Considering fusion performance, training efficiency, and memory consumption, we finally set $N = 4$.

A.10.2 Analysis of Different Reward Combinations

To evaluate the effects of different reward combinations, we conduct ablation experiments using the pixel response consistency reward (R_{pix}), structural similarity reward (R_{ssim}), and gradient detail preservation reward (R_{grad}). In the examples shown in Fig. 11, R_{pix} alone emphasizes high-intensity responses at the expense of fine details, whereas R_{ssim} alone preserves coarse structures but provides insufficient detail and target saliency. Using R_{grad} alone retains bright responses but loses some details and introduces pronounced noise near image boundaries. Combining R_{pix} and R_{ssim} produces washed-out images with substantial detail loss. The combination of R_{pix} and R_{grad} yields competitive visual quality but exhibits minor structural degradation, while combining R_{ssim} and R_{grad} preserves rich details but produces darker images.

Table 7 shows that although the full reward combination does not achieve the best result on every metric, it performs well across multiple metrics and achieves the highest Overall score of 0.889. We therefore adopt all three rewards to balance visual quality, detail preservation, and target saliency, supporting the effectiveness of our multidimensional reward design.

A.11 Model Complexity and Inference Efficiency

Table 8 reports the trainable parameter counts and FLOPs of different methods for 256×256 inputs, together with their average inference time on the 361 IR–VIS image pairs at their original resolutions in the MSRS test set. All methods are evaluated on the same hardware platform using a unified

Table 8: Trainable parameter counts and FLOPs for 256×256 inputs, and average inference time on the MSRS test set. Lower values are preferred for all metrics. The best and second-best results are highlighted in **pink** and blue, respectively. * denotes inference time measured on the CPU.

Methods	Trainable Params (M)↓	FLOPs (G)↓	Time (s/pair)↓
MetaLFusion Li et al. (2021a)	0.7453	97.2914	17.4147*
DATFuse Tang et al. (2023)	0.0106	1.1849	**0.0483**
TIMFusion Liu et al. (2024)	1.2322	46.0592	0.0548
LUT-Fuse Yi et al. (2025)	**0.0078**	**0.5169**	0.0629
SMLNet Kang et al. (2025)	0.0511	3.6522	4.7131
FreeFusion Zhao et al. (2025)	5.6551	95.3908	0.0756
DPOFusion Su et al. (2026)	795.3951	10972.4227	1.2994
DRFusion Shi et al. (2026)	31.7520	206.1793	0.0998
Ours	17.2929	1186.9761	0.3421

Table 9: Parameter counts and profiled computational costs of the proposed method for 256×256 inputs, including frozen pretrained backbones. Parameters are reported in millions (M), and FLOPs are reported in billions (G).

Module	Trainable (M)	Frozen (M)	Total (M)	FLOPs (G)
Qwen3-VL-Embedding	0.0000	2127.5320	2127.5320	1013.6416
DINOv3	0.0000	87.5665	87.5665	40.1083
Fusion modules	17.2929	0.0000	17.2929	133.2262
Entire method	**17.2929**	**2215.0985**	**2232.3914**	**1186.9761**

testing procedure. For MetaLFusion, inference time is measured on the CPU because its CUDA implementation is incompatible with the RTX 4090.

Table 9 further details the parameter counts and profiled computational costs of our method. Including the frozen DINOv3-ConvNeXt and Qwen3-VL-Embedding backbones, our method contains 2232.3914 M parameters, of which 17.2929 M are trainable, with a profiled computational cost of 1186.9761 G FLOPs. These frozen backbones still incur computational and memory costs, resulting in greater overhead than lightweight fusion methods.

Despite the use of large pretrained models, our method achieves an average inference time of 0.3421 s per image pair, covering the complete inference pipeline. This latency can be acceptable for applications that prioritize fusion quality and have modest throughput requirements. Candidate sampling is used only during training; during inference, CSD-PR remains active, while the final candidate stage decodes only the reference candidate. Although its computational and memory overhead limits real-time and resource-constrained deployment, our method primarily targets applications requiring high-quality fusion. Future work will explore model compression and more efficient semantic diagnosis to reduce computational and memory overhead while maintaining fusion quality.

A.12 Limitations and Future Work

This paper addresses two limitations of existing IR-VIS image fusion methods: their over-reliance on low-level visual constraints and their limited ability to capture semantic information and dynamic trade-offs in fusion quality. To this end, we incorporate DINOv3 and Qwen3-VL-Embedding to enhance visual representation and semantic diagnosis. These components improve both fusion quality and downstream task performance. However, introducing large-scale models increases parameter count and computational overhead, which limits deployment on resource-constrained platforms. Future work will explore more efficient semantic feedback mechanisms without explicit reliance on large-scale models. Possible directions include transferring semantic knowledge through lightweight semantic proxies or compact intermediate representations. The goal is to reduce computational cost while preserving fusion performance.