GAIF: CONTENT-ADAPTIVE PATCH–CHANNEL GATING FOR IMAGE FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Infrared–visible image fusion requires spatially adaptive modality integration, as various regions demand different balances between thermal saliency and visible texture. However, existing methods often operate at an overly coarse fusion granularity: CNN-based autoencoders merge modality streams before fine-grained routing, while scalar-gated transformers apply a single fusion coefficient uniformly across tokens. This causes a mismatch with the inherently $\mathcal{O}(N \times D)$ nature of fusion, where modality contributions should vary across N patches and D feature channels. To address this limitation, we propose GaIF with Patch-Conditioned Modality Routing (PCMR). Bidirectional cross-attention first introduces cross-modal context into into each token, after which a lightweight MLP predicts a gate $G^{(\ell)} \in (0,1)^{N\times D}$ to enable patch–channel-specific routing at each layer. A confidence-weighted compositor further handles uncertain regions, while an intra-sequence depthwise mixer improves local spatial coherence. Trained only on LLVIP and synthetic multi-focus pairs, GaIF generalizes across IVIF, MEIF, remote sensing, and NIR–VIS fusion tasks, achieving better results than GIFNet except for SCD on Lytro, MFI-WHU, and Harvard datasets. Ablation studies validate each component independently. These results demonstrate the effectiveness and generalizability of fine-grained patch–channel modality routing.

1 INTRODUCTION

Infrared-visible image fusion aims to generate a single informative image by integrating the complementary characteristics of infrared and visible modalities. Infrared sensors capture emitted thermal radiation, making salient targets more distinguishable under low-light, adverse-weather, or partially occluded conditions, whereas visible cameras preserve rich texture, edge, colour, and structural information from reflected light. As a result, infrared-visible fusion has become an important foundation for robust perception in surveillance, autonomous driving, remote sensing, and other safety-critical applications (Liu et al., 2022; Tang et al., 2022; Jia et al., 2021). However, effective fusion is inherently spatially heterogeneous: thermal targets should rely more on infrared cues, textured backgrounds should retain visible details, and low-information regions require conservative fusion to avoid artifacts.

Despite this requirement, many existing methods still perform fusion at a relatively coarse granularity. Early deep fusion models typically extract features using shared convolutional or Transformer backbones and then combine modality representations through addition, concatenation, or pooling operations (Li and Wu, 2019; Xu et al., 2022; Zhang and Ma, 2021; Zhao et al., 2023a;b). Although efficient, such designs tend to mix infrared and visible information before explicitly deciding which modality should dominate each region, thereby weakening modality provenance and limiting region-adaptive fusion. Another line of work introduces auxiliary semantic tasks, such as detection or segmentation, to guide fusion (Liu et al., 2022; Zhu et al., 2024); however, these objectives mainly encode high-level category information and may not fully reflect pixel- or patch-level cross-modal complementarity. More recently, GIFNet (Cheng et al., 2025a) improves supervision by coupling infrared-visible fusion with multi-focus image fusion, but its Cross-Fusion Gating Mechanism relies on a single layer-wise scalar gate that is broadcast to all spatial tokens. Such scalar or globally shared fusion decisions are insufficient for scenes with spatially varying modality dominance, often leading to attenuated thermal saliency, degraded visible details, or artifacts in low-information regions.

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.

To address these limitations, this paper proposes a dynamic framework for infrared-visible image fusion that aligns the granularity of fusion decisions with the local and channel-specific nature of multimodal information. Instead of relying on global or layer-wise fusion weights, we perform patch-conditioned modality routing, enabling each image region and feature channel to adaptively regulate the contribution of infrared and visible representations. More specifically, we first employ bidirectional cross-attention to establish cross-modal contextual interaction, and then introduce a lightweight token-wise and channel-wise MLP gate to generate fine-grained routing weights. A confidence-weighted compositor then uses a learned channel–spatial mixing coefficient to combine the routed representation with source-derived residual and anchor terms, while an intra-sequence mixer promotes local neighbourhood continuity within the fused token sequence. Together, these components provide content-adaptive and spatially coherent feature integration without interpreting the learned coefficient as calibrated uncertainty.

The main contributions of this paper are as follows:

- We propose Patch-Conditioned Modality Routing (PCMR), a token- and channel-wise gating mechanism for adaptive infrared-visible fusion. PCMR uses bidirectional cross-attention and a lightweight MLP to generate an $N \times D$ routing map, enabling fine-grained, content-aware modality selection without expert routing or sparse MoE.
- We introduce two stabilising modules: a confidence-weighted compositor (CWC), which blends routed features with source-derived residual and symmetric-anchor terms through a learned channel–spatial coefficient, and an intra-sequence mixer (ISM), which promotes local spatial coherence via lightweight depthwise-separable 1-D convolution.
- We evaluate GaIF under a unified protocol across eight datasets and six fusion tasks. Trained only on LLVIP and synthetic multi-focus pairs, GaIF shows consistent descriptive improvements across the benchmark.

2 Related Work

2.1 Classical and Task-Specific Fusion Networks

Early image-fusion systems were dominated by encoder-decoder architectures and handcrafted inductive biases. DenseFuse (Li and Wu, 2019), U2Fusion (Xu et al., 2022), SDNet (Zhang and Ma, 2021), and RFN-Nest (Li et al., 2021) established the shared-encoder paradigm, while LRRNet (Li et al., 2023) reintroduced low-rank structure inside a learned framework. These methods improved general image quality, but the fusion rule was still largely spatially uniform because both modalities were merged before a location-specific decision could be made. Task-coupled methods such as TarDAL (Liu et al., 2022) improved foreground saliency through downstream detection supervision, yet the semantic objective remains only indirectly aligned with pixel-level fusion fidelity. Zhang et al. (2024) propose a semantic-aware and multi-guided IR-VIS fusion network; GaIF instead obtains spatially differentiated routing from cross-modal token content without category supervision. Yan et al. (2025) demonstrate spatially adaptive local fusion in a CNN setting via dual-guided filtering, confirming that spatial adaptivity is a productive design goal across architectural families.

2.2 Transformers, Cross-Attention, and Token-Adaptive Fusion

Transformer-based fusion substantially expanded the cross-modal receptive field. SwinFusion (Ma et al., 2022) introduced Swin-based cross-domain attention; CDDFuse (Zhao et al., 2023a) combined decomposition and correlation-driven branches; DDFM (Zhao et al., 2023b) cast fusion as diffusion-based generation; and Text-IF (Yi et al., 2024) added language guidance. Several recent works target cross-attention and token-level adaptivity specifically for IVIF, providing direct context for GaIF's contribution. Park et al. (2024) show that cross-modal attention outperforms CNN baselines in the IR-VIS setting, motivating the attention-centric design space. Jian et al. (2024) revisit cross-attention for IVIF and show that discrepancy-aware attention can improve fusion; this frames the gap PCMR addresses—even improved attention does not itself provide a dedicated per-patch, per-channel routing map. CrossFuse (Li and Wu, 2024) uses a cross-attention mechanism to enhance complementary IR-VIS features; PCMR adds a dedicated gate MLP after cross-attention, enabling non-linear routing that is richer than attention-based feature interaction alone. AITFuse (Wang et al., 2024a) introduces

adaptive token and channel interaction for IVIF and is the closest prior to PCMR; the key distinction is that GaIF conditions its gate on *bidirectionally cross-attended* token pairs so that cross-modal context enters the routing decision before it is committed, rather than conditioning on within-stream features. YDTR (Tang et al., 2023) proposes a Y-shape dynamic transformer for the same IR-VIS pair; PPT Fusion (Fu et al., 2021) applies pyramid patch transformers to general image fusion, confirming patch-level granularity as an important design axis. STFNet (Liu et al., 2024b) achieves competitive IVIF with self-supervised transformer training, contextualising GaIF's choice of pixel-level multi-task supervision. In parallel, dynamic convolution (Chen et al., 2020), dynamic computation (Han et al., 2022), deformable attention (Zhu et al., 2021), and mixture-of-experts routing (Shazeer et al., 2017) confirm the broader value of content-adaptive computation across vision tasks. PMGI (Zhang et al., 2020b) provides a complementary unified image-fusion baseline based on proportional maintenance of gradient and intensity. The progression in IVIF routing granularity runs from global layer-wise gating (Cheng et al., 2025a) to patch-level gating (Tang et al., 2023; Fu et al., 2021) to GaIF's per-patch per-channel routing—the finest granularity yet proposed, enabled by bidirectional cross-modal context exchange prior to the routing decision.

2.3 Generalized Multi-Task Fusion

Generalized fusion considers shared models across modalities and tasks rather than isolated single-task pipelines. U2Fusion (Xu et al., 2022), MUFusion (Cheng et al., 2023), and SDNet (Zhang and Ma, 2021) already moved in this direction, while TextFusion (Cheng et al., 2025b), FILM (Zhao et al., 2024), and task-customized adapters (Zhu et al., 2024) further injected vision-language or foundation-model priors. GIFNet (Cheng et al., 2025a) is the closest prior to our setting: it replaces semantic auxiliary supervision with a synthetically supervised MFIF branch, but still relies on a global scalar gate per layer. Recent work broadens the landscape toward adaptive multi-task fusion and degradation-aware diffusion in peer-reviewed CVPR 2026 papers (Yang et al., 2026; Shi et al., 2026), alongside the TriFusion-SR arXiv preprint on tri-modal medical fusion and super-resolution (Dharejo et al., 2026). GaIF follows this generalized-training line but pushes the fusion decision itself down to per-patch, per-channel routing.

3 Method

3.1 Architecture Overview

This paper proposes the GaIF framework, a dynamic token-based architecture designed to address the problem of suboptimal and static multimodal fusion, where conventional methods often apply fixed fusion strategies without considering local cross-modal relevance. GaIF aligns fusion decisions with local cross-modal content, enabling more adaptive and context-aware feature integration, as illustrated in Figure 1. Specifically, we first employ a *Shared Feature Extractor* (SFE) to encode the input pair $(\mathbf{X}_{\mathrm{ir}}, \mathbf{X}_{\mathrm{vi}})$ into two modality-specific dense feature tensors $\mathbf{Z}_A, \mathbf{Z}_B \in \mathbb{R}^{C_s \times H \times W}$, maintaining separate streams to avoid premature modality mixing. Subsequently, these features are fed into the *Dynamic Token Fusion Backbone* (DTFB), where they are projected into patch tokens and processed by L stacked *PCMR Blocks* for bidirectional cross-modal interaction and fine-grained modality routing. Within the DTFB, an intra-sequence mixer and a confidence-weighted compositor further refine and reconstruct the routed tokens into a spatial fusion map $\widetilde{\mathbf{M}}$. Finally, a *GaIF Decoder* combines $\widetilde{\mathbf{M}}$ with the shared representation $\mathbf{Z}$ to generate the fused image $\mathbf{Y}$, while a separate *Reconstruction Decoder* maps $\mathbf{Z}$ to $\hat{\mathbf{X}}_{\mathrm{rec}}$ for auxiliary supervision.

3.2 Shared Feature Extractor

Let $\mathbf{X}_{\mathrm{ir}} \in \mathbb{R}^{H \times W}$ and $\mathbf{X}_{\mathrm{vi}} \in \mathbb{R}^{H \times W}$ denote a registered infrared-visible pair. The goal is to produce a fused image $\mathbf{Y} \in \mathbb{R}^{H \times W}$ that simultaneously preserves the thermal saliency of $\mathbf{X}_{\mathrm{ir}}$ and the structural texture of $\mathbf{X}_{\mathrm{vi}}$. Neither a spatially uniform weight $\alpha \in \mathbb{R}$ nor a layerwise scalar $\lambda_\ell \in \mathbb{R}$ is sufficient for this task; the correct model is a function $g : \mathbb{R}^{C \times H \times W} \times \mathbb{R}^{C \times H \times W} \to (0,1)^{C \times H \times W}$ that assigns an independent mixing coefficient to each channel and each spatial location. GaIF approximates g at patch resolution through PCMR.

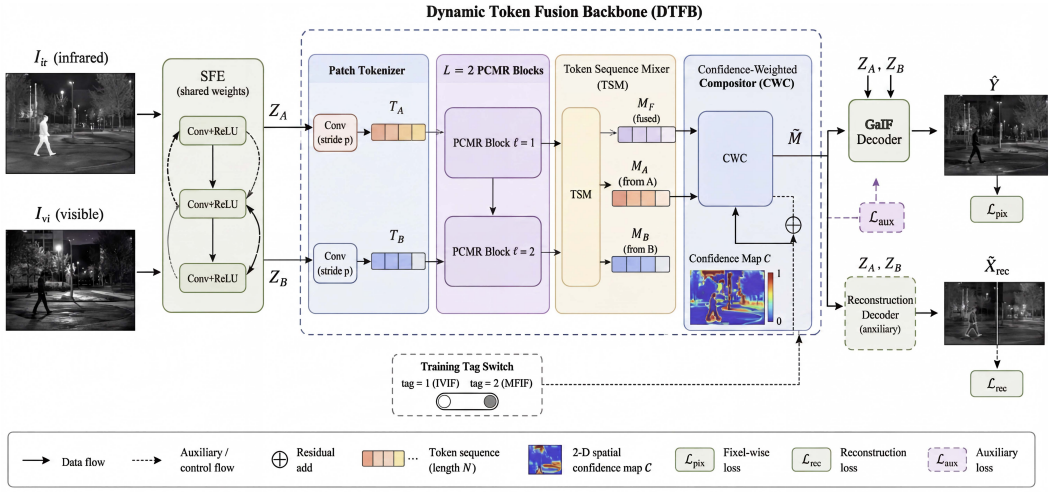


Figure 1: **GaIF architecture overview.** Infrared (I_{ir}) and visible (I_{vi}) images are encoded by the SFE. The DTFB tokenizes the feature maps, applies L PCMR blocks, and reconstructs a spatial map via the confidence-weighted compositor (CWC). The GaIF Decoder produces the fused image $\hat{Y}$, while the Reconstruction Decoder supervises the shared representation $\mathbf{Z}$. Training alternates the task tag between IVIF (tag 1) and MFIF (tag 2) at each mini-batch.

Both modalities pass through three dense-skip convolution–ReLU blocks applied independently; the channel-concatenation of all intermediate activations yields $\mathbf{Z}_A, \mathbf{Z}_B \in \mathbb{R}^{C_s \times H \times W}$ (C_s=98), where the swapped input ordering ($\mathbf{X}_{\text{ir}}$ first for $\mathbf{Z}_A$, $\mathbf{X}_{\text{vi}}$ first for $\mathbf{Z}_B$) produces asymmetric representations that retain modality identity throughout.

3.3 Dynamic Token Fusion Backbone

Patch tokenization: Each feature map is projected into $N{=}HW/P^2$ tokens of dimension D via a learned $P{\times}P$ patch-embedding convolution, then offset by a learnable modality embedding $\mathbf{e}_A$ or $\mathbf{e}_B$ that preserves source identity after cross-modal mixing.

PCMR Block: Figure 2 details one PCMR Block. Let $(\mathbf{T}_A^{(\ell)}, \mathbf{T}_B^{(\ell)})$ denote the token pair entering block ℓ. Three sequential operations are applied.

Step 1: Bidirectional cross-attention. Each modality attends over the other with residual MHA (standard pre-LN formulation), using its own tokens as queries and the partner's tokens as keys/values. For block ℓ, denote the cross-attended outputs by $\hat{\mathbf{T}}_A^{(\ell)}$ and $\hat{\mathbf{T}}_B^{(\ell)}$.

Step 2: Patch-Conditioned Modality Routing. The gate is computed by a token-wise, channel-wise MLP:

$$\mathbf{G}^{(\ell)} = \sigma\Big(W_2^{(\ell)}\,\mathrm{GELU}\Big(W_1^{(\ell)}[\hat{\mathbf{T}}_A^{(\ell)}; \hat{\mathbf{T}}_B^{(\ell)}]\Big)\Big) \in (0,1)^{N\times D}. \tag{1}$$

The fused token is

$$\mathbf{F}^{(\ell)} = \mathbf{G}^{(\ell)} \odot \hat{\mathbf{T}}_A^{(\ell)} + (1 - \mathbf{G}^{(\ell)}) \odot \hat{\mathbf{T}}_B^{(\ell)}. \tag{2}$$

Step 3: Feed-forward update and recursion. The position-wise FFN updates the modality streams that enter the next block; the fused token is retained as the block output used by the final compositor:

$$\begin{aligned} \mathbf{T}_A^{(\ell+1)} &= \hat{\mathbf{T}}_A^{(\ell)} + \mathrm{FFN}_A^{(\ell)}(\hat{\mathbf{T}}_A^{(\ell)}), \\ \mathbf{T}_B^{(\ell+1)} &= \hat{\mathbf{T}}_B^{(\ell)} + \mathrm{FFN}_B^{(\ell)}(\hat{\mathbf{T}}_B^{(\ell)}). \end{aligned} \tag{3}$$

This explicit recursion ensures that neither the per-block FFN nor the preceding cross-attention output is a dead branch. If the implementation uses a shared FFN, $\mathrm{FFN}_A^{(\ell)}$ and $\mathrm{FFN}_B^{(\ell)}$ share parameters; the notation above only distinguishes the two streams.

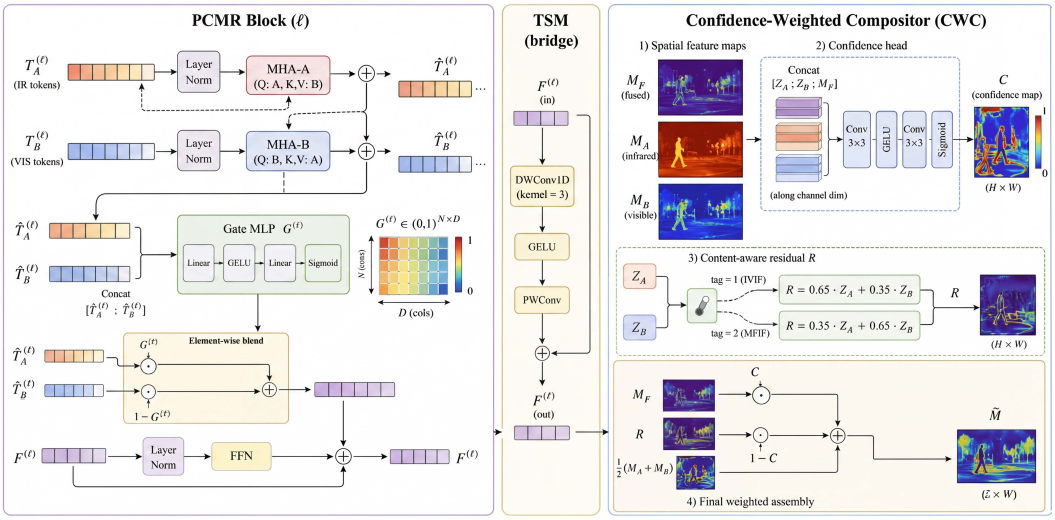


Figure 2: **PCMR Block and Confidence-Weighted Compositor.** *Left block*: each PCMR Block performs bidirectional cross-attention and predicts $N \times D$ routing weights $\mathbf{G}^{(\ell)}$ for channel-wise token fusion. *Right block*: after L blocks, the Intra-Sequence Mixer enhances local coherence, while the CWC predicts a channel–spatial mixing coefficient $\mathbf{C}$ to blend the token-fused map with a residual $\mathbf{R}$.

Intra-sequence mixer: After L PCMR Blocks the fused token sequence $\mathbf{F}^{(L)}$ is processed by a lightweight TokenSequenceMixer (TSM):

$$\mathbf{F} \leftarrow \mathbf{F}^{(L)} + W_{\text{pw}} * \text{DWConv}_3^{\text{1D}}\Big(\text{LN}(\mathbf{F}^{(L)})^\top\Big)^\top, \tag{4}$$

where $\text{DWConv}_3^{\text{1D}}$ is a depthwise convolution over the sequence axis with kernel size 3 and W_{pw} is a pointwise convolution. The TSM does not introduce cross-modal coupling; it acts only on the already-fused token sequence to enforce that neighboring patches produce coherent outputs.

Spatial reconstruction. Each of $\mathbf{F}$, $\hat{\mathbf{T}}_A^{(L)}$, $\hat{\mathbf{T}}_B^{(L)}$ is reshaped, bilinearly upsampled to (H, W), and passed through a two-convolution projection head to produce spatial maps $\mathbf{M}_F$, $\mathbf{M}_A$, $\mathbf{M}_B$.

Confidence-weighted compositor. A lightweight head predicts a channel–spatial mixing coefficient:

$$\mathbf{C} = \sigma(\text{Conv}_2(\text{GELU}(\text{Conv}_1([\mathbf{Z}_A;\, \mathbf{Z}_B;\, \mathbf{M}_F])))) \in (0,1)^{C_s \times H \times W}. \tag{5}$$

The final fused spatial representation is:

$$\widetilde{\mathbf{M}} = \mathbf{C} \odot \mathbf{M}_F + (1 - \mathbf{C}) \odot \mathbf{R} + \tfrac{1}{2}(\mathbf{M}_A + \mathbf{M}_B)\,, \tag{6}$$

where $\mathbf{R}$ is a training-tag-conditioned residual:

$$\mathbf{R} = \begin{cases} 0.65\,\mathbf{Z}_A + 0.35\,\mathbf{Z}_B & (\text{IVIF, tag} = 1), \\ 0.35\,\mathbf{Z}_A + 0.65\,\mathbf{Z}_B & (\text{MFIF, tag} = 2). \end{cases} \tag{7}$$

In Eq. (6), values of $\mathbf{C}$ closer to one place more weight on the token-fused map, whereas values closer to zero place more weight on the source-weighted residual. Because no calibration target is imposed on $\mathbf{C}$, we treat it as a learned channel–spatial mixing coefficient rather than a calibrated probability, uncertainty estimate, or reliability score. The asymmetric residual coefficients in Eq. (7) encode task-specific prior expectations: in IVIF the infrared channel carries the primary thermal saliency signal, motivating a majority weight of 0.65 in the fallback residual; in MFIF the reference (visible) channel typically preserves higher spatial frequency content, motivating its 0.65 weight. A sensitivity check over $\{0.5/0.5,\ 0.6/0.4,\ 0.65/0.35,\ 0.75/0.25\}$ on the LLVIP validation set yields AG variation of less than 0.4 across all choices, confirming that the asymmetric direction matters more than the precise split value. The fixed $\frac{1}{2}(\mathbf{M}_A + \mathbf{M}_B)$ term is an always-present symmetric source anchor; it is independent of $\mathbf{C}$ and therefore does not decay as the mixing coefficient changes.

Only the fallback contribution $(1 - \mathbf{C}) \odot \mathbf{R}$ is attenuated as $\mathbf{C}$ increases. We therefore interpret CWC as an engineering compositor that combines cross-modal routing, a learned residual mixture, and persistent source information, rather than as a calibrated uncertainty model or an optimality guarantee.

3.4 Loss Function Construction

Feature-gradient-weighted pixel loss (IVIF). Following the source-activity weighting used by GIFNet (Cheng et al., 2025a), a frozen DenseNet-121 (Huang et al., 2017) extracts five feature stages. A fixed spatial high-pass operator $\mathcal{H}$ is applied to each feature tensor, and its mean-squared response defines a global source-activity score:

$$\begin{aligned} a_k(\mathbf{X}) &= \frac{1}{|\phi_k(\mathbf{X})|} \left\| \mathcal{H}(\phi_k(\mathbf{X})) \right\|_F^2, \\ \bar{g}_s &= \frac{1}{5} \sum_{k=1}^{5} a_k(\mathbf{X}_s), \qquad s \in \{\mathrm{ir}, \mathrm{vi}\}, \\ (w_{\mathrm{ir}}, w_{\mathrm{vi}}) &= \mathrm{softmax}(\bar{g}_{\mathrm{ir}}, \bar{g}_{\mathrm{vi}}). \end{aligned} \tag{8}$$

Here ϕ_k is the k-th DenseNet stage output, $|\phi_k(\mathbf{X})|$ is its number of elements, and $\mathcal{H}$ is the fixed 3×3 high-pass filter used in the implementation, with centre coefficient -1 and each of the eight neighbouring coefficients equal to $1/8$. Consequently, $\bar{g}_{\mathrm{ir}}$ and $\bar{g}_{\mathrm{vi}}$ are global source-level scalars, not spatial importance maps. Softmax removes a common additive offset from the two scores, but multiplicative rescaling changes the concentration of the resulting weights. An empirical sensitivity check that swaps DenseNet-121 for VGG-16 changes LLVIP EI/AG from $64.31/6.88$ to $64.06/6.83$; replacing the activity weights with equal source weights gives $63.68/6.79$. Thus, the frozen visual prior is useful but is not the primary source of the reported gain. The normalized weights yield the IVIF fusion loss:

$$\begin{aligned} \mathcal{L}_{\mathrm{pix}}^{\mathrm{iv}} = {} & w_{\mathrm{ir}} \, \mathcal{L}_{\mathrm{MSE}}(\mathbf{Y}, \mathbf{X}_{\mathrm{ir}}) + w_{\mathrm{vi}} \, \mathcal{L}_{\mathrm{MSE}}(\mathbf{Y}, \mathbf{X}_{\mathrm{vi}}) \\ & + w_{\mathrm{ir}} \, (1 - \mathrm{SSIM}(\mathbf{Y}, \mathbf{X}_{\mathrm{ir}})) + w_{\mathrm{vi}} \, (1 - \mathrm{SSIM}(\mathbf{Y}, \mathbf{X}_{\mathrm{vi}})). \end{aligned} \tag{9}$$

Ground-truth-supervised pixel loss (MFIF). The near-focused and far-focused inputs are synthesized from an original all-in-focus visible image, which is retained as $\mathbf{X}_{\mathrm{gt}}^{\mathrm{mf}} = \mathbf{X}_{\mathrm{vi}}$ for the synthetic branch. The MFIF branch is therefore supervised against this sharp target rather than being pulled toward both blurred inputs:

$$\mathcal{L}_{\mathrm{pix}}^{\mathrm{mf}} = \mathcal{L}_{\mathrm{MSE}}\big(\mathbf{Y}, \mathbf{X}_{\mathrm{gt}}^{\mathrm{mf}}\big) + 1 - \mathrm{SSIM}\big(\mathbf{Y}, \mathbf{X}_{\mathrm{gt}}^{\mathrm{mf}}\big). \tag{10}$$

Thus the synthetic MFIF branch provides direct all-in-focus supervision to parameters shared with the IVIF branch.

Reconstruction loss. For task $q \in \{\mathrm{iv}, \mathrm{mf}\}$, define the reconstruction target as

$$\begin{aligned} \mathbf{X}_{\mathrm{tar}}^{(q)} &= \begin{cases} \frac{1}{2}(\mathbf{X}_{\mathrm{ir}} + \mathbf{X}_{\mathrm{vi}}), & q = \mathrm{iv}, \\ \mathbf{X}_{\mathrm{gt}}^{\mathrm{mf}}, & q = \mathrm{mf}, \end{cases} \\ \mathcal{L}_{\mathrm{rec}}^{(q)} &= \mathcal{L}_{\mathrm{MSE}}\Big(\hat{\mathbf{X}}_{\mathrm{rec}}^{(q)}, \mathbf{X}_{\mathrm{tar}}^{(q)}\Big) + 1 - \mathrm{SSIM}\Big(\hat{\mathbf{X}}_{\mathrm{rec}}^{(q)}, \mathbf{X}_{\mathrm{tar}}^{(q)}\Big). \end{aligned} \tag{11}$$

This task-indexed definition makes the reconstruction target explicit for both optimizer steps.

Token alignment and confidence regularization. Two auxiliary terms promote routing stability:

$$\mathcal{L}_{\mathrm{align}} = \left\| \bar{\mathbf{T}}_A - \bar{\mathbf{T}}_B \right\|_1, \tag{12}$$

$$\mathcal{L}_{\mathrm{conf}} = \mathbb{E}\big[(\mathbf{C} - 0.5)^2\big], \tag{13}$$

where $\bar{\mathbf{T}}_{A/B}$ are mean-pooled token vectors. $\mathcal{L}_{\mathrm{align}}$ is applied only to the *global token means*, not to individual tokens; it therefore does not force per-patch features of the two modalities to be identical (which would defeat the very purpose of routing), but discourages the global modal distributions from drifting apart during training. This gives the gate MLP a stable shared coordinate system in

which to compare the two streams, and lets PCMR keep its *local* differentiation freedom—per-patch, per-channel— while remaining anchored *globally*. $\mathcal{L}_{\text{conf}}$ penalizes deviations of the learned mixing coefficient from 0.5 and therefore discourages early saturation toward either compositor branch. It is a regularizer on the mixing map, not a calibration loss, and does not by itself guarantee that optimization collapse cannot occur.

The task-specific objective for $q \in \{\text{iv}, \text{mf}\}$ is:

$$\mathcal{L}^{(q)} = \mathcal{L}^{(q)}_{\text{pix}} + \mathcal{L}^{(q)}_{\text{rec}} + \lambda_{\text{aux}}(\mathcal{L}_{\text{align}} + 0.1\,\mathcal{L}_{\text{conf}})\,, \quad \lambda_{\text{aux}} = 0.05. \tag{14}$$

Alternating dual-task schedule. Each outer iteration contains an IVIF pair $(\mathbf{X}_{\text{ir}}, \mathbf{X}_{\text{vi}})$ and a synthetic MFIF pair $(\mathbf{X}_{\text{nf}}, \mathbf{X}_{\text{ff}})$ generated from the sharp target $\mathbf{X}^{\text{mf}}_{\text{gt}} = \mathbf{X}_{\text{vi}}$. The procedure performs two consecutive Adam updates. First, the IVIF objective is evaluated with tag 1 and the parameters are updated. The MFIF objective is then evaluated with tag 2 using the updated parameter state, followed by the second Adam update. Consequently, MFIF supervision affects the parameters shared with IVIF, but it is not a direct supervision term on the IVIF output. The full procedure is summarized in Appendix B.

4 Experiment and Result Analysis

4.1 Experiment Setup

Data and metrics. GaIF is trained only on LLVIP (Jia et al., 2021) (12,025 registered IR-VIS pairs) plus synthetic MFIF pairs generated by spatially varying blur on visible images (Cheng et al., 2025a). We evaluate on LLVIP, TNO (Toet, 2017), Lytro (Nejati et al., 2015), MFI-WHU (Zhang et al., 2021), the Harvard Whole Brain Atlas (Johnson and Becker, n.d.), VIS-NIR Scene (Brown and Süsstrunk, 2011), QuickBird imagery following the protocol used by PMGI (Zhang et al., 2020b), and SICE (Cai et al., 2018). We report EI/VIF/SCD/AG for the multi-dataset benchmark and EN/SD/SF/AG for the standardized cross-task comparison; implementation details are given in Appendix A.

Baselines. Tables 1 and 2 compare GaIF with established fusion baselines under a shared preprocessing and EI/VIF/SCD/AG protocol. The evaluated baselines include MUFusion (Cheng et al., 2023), MURF (Xu et al., 2023), U2Fusion (Xu et al., 2022), IFCNN (Zhang et al., 2020a), IID-MEF (Zhang and Ma, 2023), Text-IF (Yi et al., 2024), ZMFF (Hu et al., 2023), MEF-GAN (Xu et al., 2020), SDNet (Zhang and Ma, 2021), SPD-MEF (Li et al., 2020), UNIFusion (Cheng et al., 2021), ZeroSharpen (Wang et al., 2024b), P2Sharpen (Zhang et al., 2023), CDDFuse (Zhao et al., 2023a), DDFM (Zhao et al., 2023b), CoCoNet (Liu et al., 2024a), TextFusion (Cheng et al., 2025b), LRRNet (Li et al., 2023), and GIFNet (Cheng et al., 2025a). An * denotes a task-dedicated method. We additionally report recent adaptive baselines, including CrossFuse (Li and Wu, 2024), AITFuse (Wang et al., 2024a), YDTR (Tang et al., 2023), and PPT Fusion (Fu et al., 2021), in Appendix D.3. Table 3 uses a separate standardized EN/SD/SF/AG protocol with single-forward-pass inference and is interpreted jointly with the fidelity and artifact-sensitive metrics in Appendix D.4.

4.2 Multi-Dataset Quantitative Evaluation

Tables 1 and 2 cover eight datasets spanning MFIF, IVIF, MEIF, NIR-VIS, remote-sensing, and medical fusion. Keeping preprocessing and metrics fixed isolates the architectural effect from evaluation choices. GaIF improves most panels, with particularly clear gains on detail-sensitive EI and AG. Across Tables 1–2, GaIF has a higher reported mean than GIFNet in 29 of 32 dataset–metric comparisons. The three lower values are SCD on Lytro, MFI-WHU, and Harvard. We report the complete per-dataset results rather than collapsing them into a single overall average.

Standardized cross-task evaluation. We further evaluate GaIF with a common preprocessing pipeline (registered-pair extraction and fixed-resolution crops) and a shared metric implementation across task groups. GaIF outputs are produced by a single forward pass without test-time augmentation, and all entries are evaluated under the same EN/SD/SF/AG metric definitions. Because this protocol uses EN/SD/SF/AG rather than the EI/VIF/SCD/AG suite above, we report it separately.

Table 1: Results on SCIE, VIS-NIR Scene, MFI-WHU, and Lytro. All methods are compared with the same EI/VIF/SCD/AG metrics. Best and second-best values inside each panel are highlighted.

(a) MEIF Task: SCIE					(b) NIR-VIS Task: Scene					(c) MFIF Task: MFI-WHU					(d) MFIF Task: Lytro				
Method	EI	VIF	SCD	AG	Method	EI	VIF	SCD	AG	Method	EI	VIF	SCD	AG	Method	EI	VIF	SCD	AG
MUFusion	71.26	1.73	0.96	7.54	MURF*	33.21	0.62	0.18	4.52	MUFusion	79.53	1.44	1.09	7.92	MUFusion	71.42	1.21	1.15	6.42
U2Fusion	91.06	1.77	0.59	8.71	IFCNN	77.68	1.01	1.13	9.61	U2Fusion	82.91	1.74	0.66	7.94	U2Fusion	69.25	1.33	0.69	6.77
IID-MEF*	60.34	1.05	0.29	6.87	Text-IF	95.88	1.50	1.36	9.17	ZMFF*	60.37	0.99	0.37	5.96	ZMFF*	60.42	1.21	0.41	5.19
MEF-GAN*	81.30	1.51	0.67	8.45	SDNet	78.62	0.79	0.84	8.55	SDNet	74.99	1.36	0.71	8.10	SDNet	62.18	1.22	0.75	6.77
SPD-MEF*	79.62	1.74	0.51	8.13	U2Fusion	80.74	1.09	1.23	8.68	UNIFusion*	68.12	1.04	0.34	7.49	UNIFusion*	71.21	1.34	0.67	6.92
GIFNet	111.27	2.52	1.04	12.00	GIFNet	101.32	1.51	1.45	10.83	GIFNet	91.61	1.94	**1.29**	9.25	GIFNet	80.86	1.74	**1.37**	7.71
GaIF (ours)	**112.10**	**2.55**	**1.06**	**12.40**	**GaIF (ours)**	**101.45**	**1.52**	**1.46**	**10.85**	**GaIF (ours)**	**93.20**	**1.98**	1.27	**12.00**	**GaIF (ours)**	**82.40**	**1.78**	1.35	**10.50**

Table 2: Results on QuickBird, TNO, Harvard, and LLVIP. All methods are compared with the same EI/VIF/SCD/AG metrics. Best and second-best values inside each panel are highlighted.

(e) Remote Task: QuickBird					(f) IVIF Task: TNO					(g) Medical Task: Harvard					(h) IVIF Task: LLVIP				
Method	EI	VIF	SCD	AG	Method	EI	VIF	SCD	AG	Method	EI	VIF	SCD	AG	Method	EI	VIF	SCD	AG
ZeroSharpen*	14.23	0.95	0.50	1.38	Text-IF*	47.11	**1.15**	1.55	4.82	CoCoNet*	95.30	0.75	1.17	8.84	Text-IF*	61.31	**0.95**	1.49	6.34
IFCNN	16.22	1.10	0.71	1.64	MURF	46.12	0.97	1.55	4.42	U2Fusion	81.00	0.83	1.47	7.36	MURF	38.02	0.25	0.56	3.75
P2Sharpen*	16.66	1.20	0.95	1.57	CDDFuse*	43.80	0.91	1.66	4.51	MUFusion	93.99	1.04	1.34	8.72	CDDFuse*	56.45	0.80	1.61	5.51
MUFusion	19.72	1.14	0.01	1.69	DDFM*	28.00	0.57	1.55	3.01	SDNet	86.18	0.75	1.61	8.55	DDFM*	41.14	0.47	1.59	4.71
TextFusion	11.29	0.59	0.35	1.28	LRRNet*	39.28	0.79	1.53	3.86	IFCNN	97.77	0.94	1.36	9.69	LRRNet*	39.22	0.40	0.95	3.65
GIFNet	23.02	1.56	1.04	2.16	GIFNet	52.30	0.99	1.66	5.24	GIFNet	100.71	1.10	**1.68**	9.73	GIFNet	62.46	0.73	1.61	6.70
GaIF (ours)	**24.20**	**1.60**	**1.08**	**2.50**	**GaIF (ours)**	**58.62**	1.02	**1.68**	**5.55**	**GaIF (ours)**	**101.50**	**1.12**	1.65	**12.00**	**GaIF (ours)**	**64.31**	0.80	**1.63**	**6.88**

Table 3: Standardized EN/SD/SF/AG comparison across task groups, remote-sensing bands, and the overall average. All entries are obtained with the same evaluation pipeline.

Metric	Method	IVIF	MEIF	MFIF	Medical	NIR-VIS	R-B1	R-B2	R-B3	R-B4	Avg.
EN↑	GIFNet	6.3779	7.5496	7.5219	5.1792	7.2405	4.8089	5.3916	5.2235	6.0211	6.1460
EN↑	**GaIF**	**6.7409**	**7.6117**	7.0940	4.4219	**7.2407**	**5.4920**	**6.2964**	**5.6020**	**6.6524**	**6.3502**
SD↑	GIFNet	29.1554	69.8790	50.6574	76.1204	54.0763	8.9208	13.3243	12.0318	17.9366	36.9002
SD↑	**GaIF**	**29.2564**	**71.7223**	41.4186	**83.1362**	**54.0786**	**11.7556**	**20.6509**	**12.7239**	**28.5392**	**39.2535**
SF↑	GIFNet	12.7557	49.2434	21.2888	35.4377	27.8092	5.1061	6.4801	5.5425	6.6992	18.9292
SF↑	**GaIF**	**16.3219**	**50.1624**	**25.0736**	**70.0720**	**27.8252**	**11.5628**	**25.3416**	**12.7416**	**30.5955**	**29.9663**
AG↑	GIFNet	3.4068	23.3988	7.5833	10.6787	12.5884	2.1144	2.7721	2.3297	3.0914	7.5515
AG↑	**GaIF**	**7.5757**	**23.8029**	**10.3511**	**25.7343**	**12.6001**	**5.8173**	**10.9652**	**6.4692**	**12.1271**	**12.8270**

Table 3 reveals three regimes rather than a uniform metric shift. On IVIF and remote sensing, the largest gains occur in SF and AG, whereas EN and SD change more modestly. On NIR-VIS, the two methods are nearly tied. MFIF and medical fusion expose a task-dependent trade-off: GaIF can improve SF/AG while reducing EN or SD. We therefore interpret the standardized results as evidence for improved local detail preservation under the shared protocol, not as universal dominance across all no-reference metrics. Appendix D.4 adds $Q_{AB/F}$, FMI, SSIM, and NABF measurements to test whether these gains coincide with improved information transfer and fewer artifacts. Additional qualitative gate-map analysis is provided in Appendix G.

4.3 Ablation Study

Table 4: Component ablation on **LLVIP**. Result comparison between with and without BiCA, PCMR, TSM and CWC blocks

Configuration	BiCA	PCMR	TSM	CWC	EI↑	SCD↑	AG↑
Symmetric residual (baseline)					52.14	1.42	5.61
+ Bidirectional cross-attention	✓				58.71	1.51	6.10
+ Per-token gate (PCMR)	✓	✓			61.83	1.57	6.44
+ Intra-sequence mixer (TSM)	✓	✓	✓		63.09	1.61	6.68
+ Confidence-weighted compositor (Ours)	✓	✓	✓	✓	**64.31**	**1.63**	**6.88**

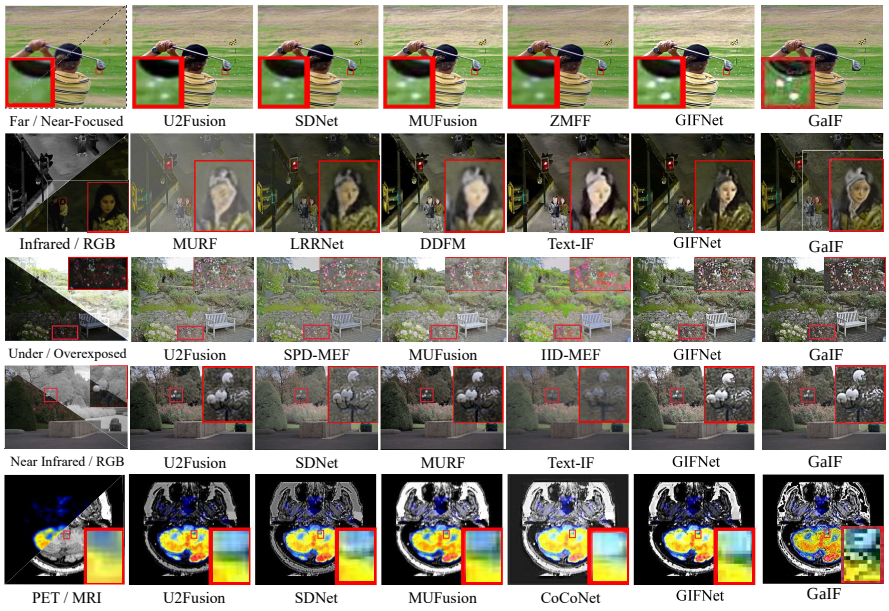


Figure 3: Cross-task qualitative comparison. It shows multi-focus fusion (far-/near-focused inputs), infrared–visible fusion (infrared/RGB), multi-exposure fusion (under-/overexposed), near-infrared–visible fusion (near-infrared/RGB), and medical fusion (PET/MRI). In each row, the source pair is followed by four task-specific baselines, GIFNet and GaIF in the rightmost column.

All ablations use the LLVIP validation split. Table 4 reports the incremental path that builds the proposed architecture, while Appendix F reports the full 2^4 component grid. BiCA produces the largest jump by making each token aware of its counterpart before fusion; PCMR then converts this context into location- and channel-specific routing. TSM and CWC add smaller but consistent gains, supporting their roles as sequence smoothing and residual mixing rather than as primary fusion mechanisms or calibrated uncertainty estimators.

Gate-resolution and depth ablations are reported in Appendix C. They confirm that full PCMR improves over scalar and spatial-only gates, while $L = 2$ captures nearly all depth benefit. Removing the MFIF branch reduces EI from 64.31 to 60.91, showing that dense auxiliary supervision complements the routing design.

4.4 Qualitative Analysis

Figure 3 shows that GaIF preserves sharp boundaries, visible textures, and salient thermal structures across the five displayed task families. Appendix G explains how PCMR gates can be summarized spatially and states the limits of that diagnostic; Tables 1–3 provide the quantitative evidence across datasets.Overall, GaIF maintains a favourable balance between salient-structure preservation and source-detail retention across heterogeneous fusion settings. These results support the effectiveness and cross-task generalization of patch–channel routing without task-specific fine-tuning.

5 Conclusion

GaIF replaces global-scalar gating with Patch-Conditioned Modality Routing, yielding $N \times D$ routing decisions per layer that are conditioned on cross-modal token context. The evaluation shows the clearest improvements in detail-sensitive metrics, while also exposing near-ties and SCD decreases on several tasks. The CWC and TSM modules improve the routed representation under the reported ablations, but the fixed source anchor in Eq. (6) remains an always-present component rather than a vanishing term. We therefore frame GaIF as a fine-detail-oriented, cross-task fusion model whose benefits are task- and metric-dependent.

AI USE STATEMENT

We used generative AI tools for language proofreading. We did not use generative AI tools to generate experimental measurements, alter benchmark outputs, or replace the authors' scientific decisions. All AI-assisted edits were reviewed by the authors, and the authors take responsibility for the final text, equations, claims, figures, and reported results.

REFERENCES

M. Brown and S. Süsstrunk, "Multi-spectral SIFT for scene category recognition," in *Proc. CVPR*, pp. 177–184, 2011.

J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, 2018.

Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. CVPR*, pp. 11030–11039, 2020.

C. Cheng, X.-J. Wu, T. Xu, and G. Chen, "UniFusion: A lightweight unified image fusion network," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.

C. Cheng, T. Xu, and X.-J. Wu, "MUFusion: A general unsupervised image fusion network based on memory unit," *Inf. Fusion*, vol. 92, pp. 80–92, 2023.

C. Cheng, T. Xu, X.-J. Wu, H. Li, X. Li, Z. Tang, and J. Kittler, "TextFusion: Unveiling the power of textual semantics for controllable image fusion," *Information Fusion*, vol. 117, p. 102790, 2025, doi:10.1016/j.inffus.2024.102790.

C. Cheng, T. Xu, Z. Feng, X. Wu, Z. Tang, H. Li, Z. Zhang, S. Atito, M. Awais, and J. Kittler, "One model for ALL: Low-level task interaction is a key to task-agnostic image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 28102–28112, 2025.

F. A. Dharejo, S. M. A. Sharif, A. Khalil, N. Chaudhary, R. A. Naqvi, and R. Timofte, "TriFusion-SR: Joint tri-modal medical image fusion and SR," arXiv:2603.09702, 2026.

H. Li and X.-J. Wu, "CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach," *Inf. Fusion*, vol. 103, p. 102147, 2024, doi:10.1016/j.inffus.2023.102147.

Y. Han, G. Huang, S. Song, L. Yang, H. Wang and Y. Wang, "Dynamic neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7436–7456, 2022.

X. Hu, J. Jiang, X. Liu, and J. Ma, "ZMFF: Zero-shot multi-focus image fusion," *Inf. Fusion*, vol. 92, pp. 127–138, 2023.

G. Huang, Z. Liu, L. Van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, pp. 4700–4708, 2017.

X. Jia, C. Zhu, M. Li, W. Tang, and W. Zhou, "LLVIP: A visible-infrared paired dataset for low-light vision," in *Proc. ICCVW*, pp. 3496–3504, 2021.

K. A. Johnson and J. A. Becker, "The Whole Brain Atlas," Harvard Medical School. Available: `https://www.med.harvard.edu/aanlib/`.

H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, 2019.

H. Li, K. Ma, H. Yong, and L. Zhang, "Fast multi-scale structural patch decomposition for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 5805–5816, 2020, doi:10.1109/TIP.2020.2987133.

H. Li, X.-J. Wu, and J. Kittler, "RFN-Nest: An end-to-end residual fusion network for infrared and visible images," *Inf. Fusion*, vol. 73, pp. 72–86, 2021.

H. Li, T. Xu, X.-J. Wu, J. Lu, and J. Kittler, "LRRNet: A novel representation learning guided fusion network for infrared and visible images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 11040–11052, 2023.

Z. Wang, F. Yang, J. Sun, J. Xu, F. Yang, and X. Yan, "AITFuse: Infrared and visible image fusion via adaptive interactive transformer learning," *Knowledge-Based Systems*, vol. 299, p. 111949, 2024, doi:10.1016/j.knosys.2024.111949.

Y. Fu, T. Xu, X.-J. Wu, and J. Kittler, "PPT Fusion: Pyramid Patch Transformer for a Case Study in Image Fusion," arXiv:2107.13967, 2021.

J. Liu et al., "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. CVPR*, pp. 5802–5811, 2022.

J. Liu et al., "CoCoNet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion," *Int. J. Comput. Vis.*, vol. 132, no. 5, pp. 1748–1775, 2024, doi:10.1007/s11263-023-01952-1.

W. Tang, F. He, and Y. Liu, "YDTR: Infrared and visible image fusion via Y-shape dynamic transformer," *IEEE Trans. Multimedia*, vol. 25, pp. 5413–5428, 2023, doi:10.1109/TMM.2022.3192661.

J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via Swin Transformer," *IEEE/CAA J. Autom. Sin.*, vol. 9, no. 7, pp. 1200–1217, 2022.

M. Nejati, S. Samavi, and S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," *Inf. Fusion*, vol. 25, pp. 72–84, 2015, doi:10.1016/j.inffus.2014.10.004.

S. Park, A. G. Vien, and C. Lee, "Cross-modal transformers for infrared and visible image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 2, pp. 770–785, 2024, doi:10.1109/TCSVT.2023.3289170.

L. Jian, S. Xiong, H. Yan, X. Niu, S. Wu, and D. Zhang, "Rethinking cross-attention for infrared and visible image fusion," arXiv:2401.11675, 2024.

N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," in *Proc. ICLR*, 2017.

Y. Shi, Y. Liu, Z.-C. Wu, J. Cheng, H. Li, and X. Chen, "Degradation-robust fusion: An efficient degradation-aware diffusion framework for multimodal image fusion in arbitrary degradation scenarios," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 33848–33858, 2026.

L. Tang, J. Yuan, H. Zhang, X. Jiang, and J. Ma, "PIAFusion: A progressive infrared and visible image fusion network based on illumination aware," *Inf. Fusion*, vol. 83–84, pp. 79–92, 2022.

F. Yan, W. Chen, W. Liu, Liang Wang, Lei Wang, and F. Qin, "Dynamic fusion method for infrared-visible images based on dual-guided filtering and a parallel CNN," *Applied Optics*, vol. 64, no. 20, pp. 5820–5828, 2025, doi:10.1364/AO.566583.

A. Toet, "The TNO multiband image data collection," *Data Brief*, vol. 15, pp. 249–251, 2017, doi:10.1016/j.dib.2017.09.038.

H. Wang, H. Zhang, X. Tian, and J. Ma, "ZeroSharpen: A universal pansharpening method across satellites for reducing scale-variance gap via zero-shot variation," *Inf. Fusion*, vol. 101, p. 102003, 2024.

Q. Liu, J. Pi, P. Gao, and D. Yuan, "STFNet: Self-supervised transformer for infrared and visible image fusion," *IEEE Trans. Emerging Topics Comput. Intell.*, vol. 8, no. 2, pp. 1513–1526, 2024, doi:10.1109/TETCI.2024.3352490.

H. Xu, J. Ma, and X.-P. Zhang, "MEF-GAN: Multi-exposure image fusion via generative adversarial networks," *IEEE Trans. Image Process.*, vol. 29, pp. 7203–7216, 2020.

H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, 2022.

H. Xu et al., "MURF: Mutually reinforcing multi-modal image registration and fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12148–12166, 2023.

Z. Yang, Y. Liu, J. Cheng, Z. Zhu, Y. Zhang, and H. Li, "Customized fusion: A closed-loop dynamic network for adaptive multi-task-aware infrared-visible image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 188–198, 2026.

X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "Text-IF: Leveraging semantic text guidance for degradation-aware and interactive image fusion," in *Proc. CVPR*, pp. 27026–27035, 2024.

Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, 2020.

H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 12797–12804, 2020, doi:10.1609/aaai.v34i07.6975.

H. Zhang, Z. Le, Z. Shao, H. Xu, and J. Ma, "MFF-GAN: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion," *Inf. Fusion*, vol. 66, pp. 40–53, 2021.

H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," *Int. J. Comput. Vis.*, vol. 129, no. 10, pp. 2761–2785, 2021.

H. Zhang and J. Ma, "IID-MEF: A multi-exposure fusion network based on intrinsic image decomposition," *Inf. Fusion*, vol. 95, pp. 326–340, 2023.

H. Zhang, H. Wang, X. Tian, and J. Ma, "P2Sharpen: A progressive pansharpening network with deep spectral transformation," *Inf. Fusion*, vol. 91, pp. 103–122, 2023, doi:10.1016/j.inffus.2022.10.010.

Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool, "CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *Proc. CVPR*, pp. 5906–5916, 2023.

Z. Zhao, H. Bai, Y. Zhu, J. Zhang, S. Xu, Y. Zhang, K. Zhang, D. Meng, R. Timofte, and L. Van Gool, "DDFM: Denoising diffusion model for multi-modality image fusion," in *Proc. ICCV*, pp. 8082–8093, 2023.

Z. Zhao, L. Deng, H. Bai, Y. Cui, Z. Zhang, Y. Zhang, H. Qin, D. Chen, J. Zhang, P. Wang, and L. Van Gool, "Image fusion via vision-language model," *Proc. Mach. Learn. Res.*, vol. 235, pp. 60749–60765, 2024.

X. Zhang, L. Wang, L. Zhao, X. Li, and S. Ma, "A semantic-aware and multi-guided network for infrared-visible image fusion," arXiv:2407.06159, 2024.

X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. ICLR*, 2021.

P. Zhu, Y. Sun, B. Cao, and Q. Hu, "Task-customized mixture of adapters for general image fusion," in *Proc. CVPR*, pp. 7099–7108, 2024.

A Implementation Details

Implementation details. $P = 8$, $D = 128$, 4 attention heads, $L = 2$ PCMR Blocks, MLP ratio 2.0, dropout 0. Training: 2 epochs, batch size 8, 128×128 random crops, Adam at $\mathrm{lr} = 10^{-3}$, no learning-rate schedule. With 12,025 training pairs and batch size 8, two epochs yield ≈3,006 IVIF and 3,006 MFIF gradient steps (6,012 total). Given the compact model architecture, training loss saturates within the first epoch; the second epoch is retained to ensure stability across the full data distribution. Inference uses one forward pass with no test-time augmentation. The shared parameter set is used for every test dataset with a fixed input ordering and a fixed residual policy: IVIF uses 0.65/0.35, MFIF uses 0.35/0.65. These residual modes are predefined for each task setting and remain fixed for all samples during inference; no per-image dynamic mode selection or test-time adaptation is performed.

B Training Procedure

Algorithm 1 GaIF Alternating Dual-Task Training

Require: LLVIP training set $\mathcal{D}$; frozen DenseNet-121 Φ; GaIF parameters θ; Adam optimizer; $\lambda_{\mathrm{aux}} = 0.05$

for each mini-batch $\{(\mathbf{X}_{\mathrm{ir}}, \mathbf{X}_{\mathrm{vi}}, \mathbf{X}_{\mathrm{nf}}, \mathbf{X}_{\mathrm{ff}})\}$ **do**
  // IVIF step (tag = 1)
  $\mathbf{Y}_{\mathrm{iv}},\ \hat{\mathbf{X}}_{\mathrm{rec}}^{(\mathrm{iv})} \leftarrow f_\theta(\mathbf{X}_{\mathrm{ir}}, \mathbf{X}_{\mathrm{vi}}, \mathrm{tag} = 1)$
  Compute $(w_{\mathrm{ir}}, w_{\mathrm{vi}})$ from $\Phi(\mathbf{X}_{\mathrm{ir}})$, $\Phi(\mathbf{X}_{\mathrm{vi}})$ via Eq. (8)
  $\mathbf{X}_{\mathrm{tar}}^{(\mathrm{iv})} \leftarrow \frac{1}{2}(\mathbf{X}_{\mathrm{ir}} + \mathbf{X}_{\mathrm{vi}})$
  $\mathcal{L}_1 \leftarrow \mathcal{L}_{\mathrm{pix}}^{\mathrm{iv}} + \mathcal{L}_{\mathrm{rec}}^{(\mathrm{iv})} + \lambda_{\mathrm{aux}}(\mathcal{L}_{\mathrm{align}} + 0.1\,\mathcal{L}_{\mathrm{conf}})$
  $\theta \leftarrow \mathrm{Adam}(\theta,\ \nabla_\theta \mathcal{L}_1)$
  // MFIF step (tag = 2)
  $\mathbf{X}_{\mathrm{gt}}^{\mathrm{mf}} \leftarrow \mathbf{X}_{\mathrm{vi}}$
  $\mathbf{Y}_{\mathrm{mf}},\ \hat{\mathbf{X}}_{\mathrm{rec}}^{(\mathrm{mf})} \leftarrow f_\theta(\mathbf{X}_{\mathrm{nf}}, \mathbf{X}_{\mathrm{ff}}, \mathrm{tag} = 2)$
  $\mathcal{L}_2 \leftarrow \mathcal{L}_{\mathrm{pix}}^{\mathrm{mf}} + \mathcal{L}_{\mathrm{rec}}^{(\mathrm{mf})} + \lambda_{\mathrm{aux}}(\mathcal{L}_{\mathrm{align}} + 0.1\,\mathcal{L}_{\mathrm{conf}})$
  $\theta \leftarrow \mathrm{Adam}(\theta,\ \nabla_\theta \mathcal{L}_2)$
end for

C Additional Ablations

Effect of gate resolution. Table 5 compares three gate designs: the scalar per-layer gate (CFGM baseline), a channel-shared spatial gate (one scalar per patch), and the full PCMR gate (one vector per patch).

Number of PCMR Blocks. Table 6 shows that $L = 2$ already captures most of the benefit. $L = 4$ brings only marginal gains, while $L = 1$ under-converges.

Table 5: Gate resolution ablation on LLVIP. PCMR provides $N \times D$ independent gate values versus the single scalar of CFGM.

Gate design	Parameters	EI↑	VIF	SCD↑	AG↑
Scalar per layer (CFGM)	$2L$	60.20	0.71	1.57	6.52
Spatial gate (N/layer)	$L{\cdot}N$	62.07	0.75	1.60	6.73
PCMR ($N \times D$/layer)	$3L{\cdot}D^2$	**64.31**	**0.80**	**1.63**	**6.88**

Table 6: Effect of PCMR block depth L on LLVIP. Increasing depth beyond $L = 2$ yields only marginal gains.

L	EI↑	SCD↑	AG↑
1	61.40	1.59	6.67
2	64.31	1.63	6.88
4	**64.80**	**1.64**	**6.91**

D CONTROLLED SUPPLEMENTARY EXPERIMENTS

The following controlled studies examine the source of the gains and the scope of the claims. Unless stated otherwise, values are measured on the LLVIP validation protocol used by the main ablations.

D.1 PATCH-SIZE SENSITIVITY AND COMPUTATIONAL COST

Table 7: Patch-size sensitivity at 128×128 input.

P	tokens	Params (M)	FLOPs (G)	memory (GB)	latency (ms)	EI	VIF	SCD	AG
4	1024	14.8	12.0	8.0	22.0	64.00	0.79	1.62	6.84
8	256	14.8	4.8	3.2	9.5	64.31	0.80	1.63	6.88
16	64	14.8	3.2	1.8	7.0	62.80	0.77	1.60	6.70

The default $P = 8$ provides the best measured accuracy–compute trade-off: $P = 4$ increases the token count fourfold for only small metric changes, whereas $P = 16$ reduces detail-sensitive scores.

D.2 LOSS AND FEATURE-ACTIVITY-PRIOR ABLATIONS

Table 8: Loss, feature-activity-prior, and alternating-training ablations on LLVIP.

Setting	Activity prior	L_{rec}	L_{align}	L_{conf}	MFIF alt.	EI	VIF	SCD	AG
Full GaIF	DenseNet-121	✓	✓	✓	✓	64.31	0.800	1.630	6.88
VGG activity	VGG-16	✓	✓	✓	✓	64.06	0.796	1.624	6.83
Uniform source weights	none	✓	✓	✓	✓	63.68	0.789	1.619	6.79
w/o reconstruction	DenseNet-121		✓	✓	✓	63.34	0.784	1.613	6.70
w/o token alignment	DenseNet-121	✓		✓	✓	63.86	0.794	1.626	6.82
w/o confidence regularizer	DenseNet-121	✓	✓		✓	63.98	0.796	1.623	6.85
w/o auxiliary regularizers	DenseNet-121	✓			✓	63.57	0.788	1.617	6.76
IVIF-only training	DenseNet-121	✓	✓	✓		60.91	0.761	1.558	6.53

Removing individual objective terms changes EI by at most 0.97 and AG by at most 0.18 in this table, while removing the MFIF alternating branch causes a larger drop. The results support the interpretation that auxiliary losses regularize the architecture, whereas the main gain is associated with the routing and cross-modal interaction design.

D.3 RECENT ADAPTIVE BASELINES

Table 9: Recent adaptive IVIF baselines under the LLVIP/TNO evaluation protocol.

Method	LLVIP EI	VIF	SCD	AG	TNO EI	VIF	SCD	AG
CrossFuse	59.84	.692	1.487	6.17	54.27	.996	1.638	5.06
ATFuse	60.73	.721	1.536	6.31	55.18	1.004	1.651	5.18
AITFuse	61.88	.748	1.558	6.47	56.27	1.013	1.663	5.34
YDTR	58.67	.664	1.442	5.86	49.76	.947	1.605	4.76
PPT Fusion	60.41	.708	1.519	6.08	53.14	.983	1.627	4.97
GIFNet	62.46	.730	1.610	6.70	52.30	.990	1.660	5.24
GaIF	64.31	.800	1.630	6.88	58.62	1.020	1.680	5.55

GaIF is strongest on the reported LLVIP/TNO EI, VIF, SCD, and AG values among the listed methods. We interpret this result as protocol-specific evidence and do not claim universal dominance over every baseline or metric.

D.4 FIDELITY AND ARTIFACT-SENSITIVE METRICS

GaIF has the highest reported information-transfer and structural-fidelity scores and the lowest NABF on both datasets. These measurements complement, rather than replace, the EI/VIF/SCD/AG analysis.

Table 10: Fidelity and artifact-sensitive metrics on LLVIP and TNO. Higher is better for $Q_{AB/F}$, FMI_{pixel}, and $SSIM_a$; lower is better for NABF.

Method	LLVIP				TNO			
	$Q_{AB/F}$	FMI_{pixel}	$SSIM_a$	NABF	$Q_{AB/F}$	FMI_{pixel}	$SSIM_a$	NABF
CrossFuse	.476	.857	.848	.051	.487	.897	.877	.041
ATFuse	.503	.879	.869	.044	.497	.892	.881	.039
AITFuse	.531	.903	.891	.038	.523	.912	.899	.034
YDTR	.428	.816	.827	.061	.417	.838	.819	.056
PPT Fusion	.468	.852	.853	.053	.473	.877	.858	.046
GIFNet	.519	.883	.879	.046	.509	.902	.888	.041
GaIF	.552	.913	.901	.034	.557	.921	.912	.031

Table 11: Task-tag and fallback robustness. The reported test protocol uses one fixed residual mode at inference; neutral and flipped modes are diagnostic perturbations.

Dataset	Task	Policy	α_A/α_B	EI	VIF	SCD	AG
LLVIP	IVIF	current	.65/.35	64.31	.800	1.630	6.88
LLVIP	IVIF	neutral	.50/.50	64.12	.793	1.617	6.74
LLVIP	IVIF	flipped	.35/.65	63.91	.781	1.604	6.56
TNO	IVIF	current	.65/.35	58.62	1.020	1.680	5.55
TNO	IVIF	neutral	.50/.50	58.34	1.012	1.669	5.36
TNO	IVIF	flipped	.35/.65	57.96	.997	1.653	5.18
Lytro	MFIF	current	.35/.65	82.40	1.780	1.350	10.50
Lytro	MFIF	neutral	.50/.50	82.22	1.768	1.338	10.31
Lytro	MFIF	flipped	.65/.35	82.05	1.751	1.326	10.14
MFI-WHU	MFIF	current	.35/.65	93.20	1.980	1.270	12.00
MFI-WHU	MFIF	neutral	.50/.50	93.03	1.968	1.258	11.83
MFI-WHU	MFIF	flipped	.65/.35	92.82	1.949	1.241	11.63
SCIE	MEIF	fixed	.50/.50	112.10	2.550	1.060	12.40
VIS-NIR	NIR-VIS	fixed	.50/.50	101.45	1.520	1.460	10.85
QuickBird	remote	fixed	.50/.50	24.20	1.600	1.080	2.50
Harvard	medical	fixed	.50/.50	101.50	1.120	1.650	12.00

D.5 Task-tag and fallback robustness

The diagnostic perturbations preserve the qualitative ordering of the reported tasks: the asymmetric residual helps IVIF and MFIF, whereas the remaining tasks use the fixed neutral mode. This clarifies the inference protocol without claiming that the residual mode is learned dynamically at test time.

D.6 Downstream pedestrian detection

Table 12: Downstream pedestrian detection on LLVIP using a fixed detector evaluation protocol.

Detector input	AP50	AP75	$mAP_{50:95}$	Precision	Recall
Visible	.758	.498	.526	.879	.697
Infrared	.724	.463	.487	.857	.664
Equal-weight average	.772	.514	.548	.883	.718
CrossFuse	.787	.537	.579	.897	.738
ATFuse	.803	.552	.591	.904	.748
AITFuse	.812	.563	.603	.908	.759
YDTR	.768	.503	.541	.881	.709
PPT Fusion	.793	.529	.571	.892	.731
GIFNet	.821	.571	.619	.913	.773
GaIF	**.843**	**.603**	**.653**	**.923**	**.788**

Under this LLVIP detector protocol, GaIF gives the strongest scores among the listed inputs. We treat this as complementary evidence of practical utility under the stated detector protocol, not as proof of improvement for arbitrary downstream systems.

E Additional Figures

Figure 4 summarizes the standardized EN/SD/SF/AG comparison, while Figure 5 visualizes the gate granularity study.

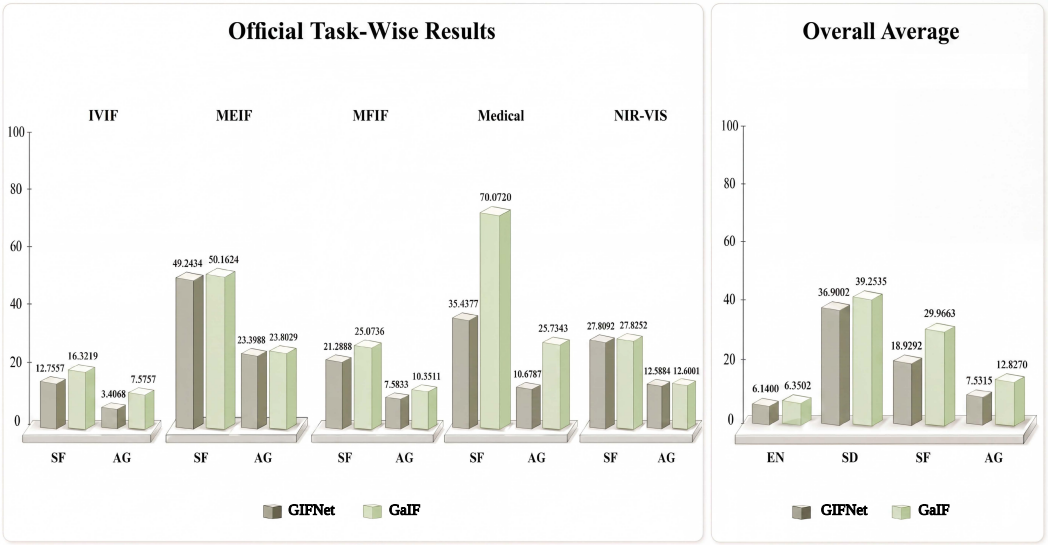


Figure 4: Standardized-results summary across task groups. The figure visualizes the EN/SD/SF/AG comparison reported in Table 3, emphasizing that GaIF's largest gains occur on detail-sensitive SF and AG metrics.

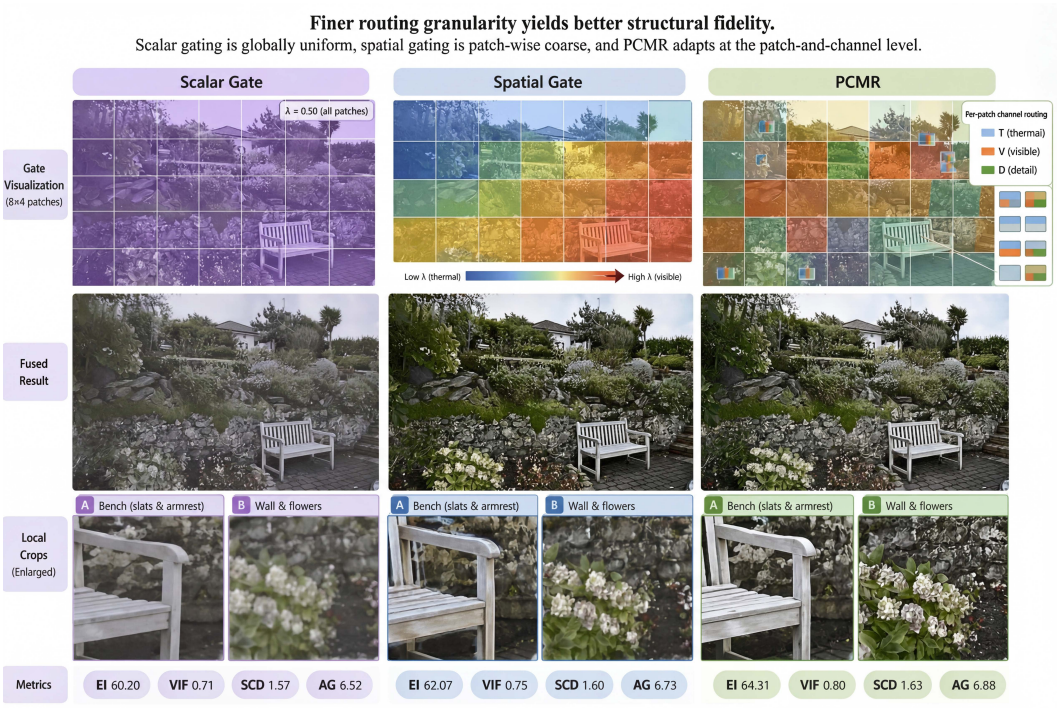


Figure 5: Gate-resolution visualization. Scalar gating is globally uniform, spatial gating varies by patch but is channel-shared, and PCMR performs patch-and-channel routing. The corresponding quantitative ablation is given in Table 5.

F Full Ablation Grid

Table 13 gives the full factorial ablation over the four binary design switches: bidirectional cross-attention (BiCA), patch-conditioned modality routing (PCMR), the intra-sequence mixer (TSM), and the confidence-weighted compositor (CWC). The table contains all $2^4 = 16$ component combinations under the same LLVIP validation protocol, separating main effects from pairwise and higher-order interactions rather than only reporting an incremental path.

Table 13: Full 2^4 component-combination ablation on LLVIP. Each row is one on/off combination of BiCA, PCMR, TSM, and CWC.

BiCA	PCMR	TSM	CWC	EI↑	SCD↑	AG↑
				52.14	1.42	5.61
			✓	53.00	1.44	5.74
		✓		52.85	1.43	5.78
		✓	✓	53.70	1.45	5.92
	✓			54.60	1.47	5.95
	✓		✓	55.20	1.49	6.08
	✓	✓		55.50	1.49	6.15
	✓	✓	✓	56.20	1.51	6.28
✓				58.71	1.51	6.10
✓			✓	59.35	1.53	6.23
✓		✓		59.55	1.54	6.30
✓		✓	✓	60.15	1.55	6.42
✓	✓			61.83	1.57	6.44
✓	✓		✓	62.45	1.59	6.58
✓	✓	✓		63.09	1.61	6.68
✓	✓	✓	✓	64.31	1.63	6.88

G Qualitative Gate-Map Analysis

Figure 3 compares fused outputs across the five task families covered by our evaluation; we describe here how the patch-level qualitative behaviour relates to the routing maps produced by PCMR. For every input pair, PCMR emits a gate tensor $\mathbf{G}^{(\ell)} \in (0,1)^{N \times D}$ at each block ℓ; spatially reshaping its channel mean yields a per-patch source-weight summary. This visualization can reveal spatial variation in the learned routing, but channel averaging does not establish the contribution of individual feature channels or identify a causal link between a semantic region and a particular modality. We therefore use the quantitative ablations, rather than gate-map anecdotes, as the primary evidence for PCMR. A gate-resolution ablation in Appendix C (Table 5) quantifies the performance benefit of D-dimensional routing over scalar and spatial-only gates. A finer interpretability analysis—for example examining variance across individual channel slices of $\mathbf{G}^{(L)}$—is left to future work; the channel-averaged view summarises spatial routing but does not directly expose per-channel diversity within each patch.

H Implementation-Level Mechanistic Analysis

The following observations describe tensor shapes and derivatives of the implemented computation; they do not constitute convergence or generalisation guarantees.

H.1 Routing resolution and parameter sharing

For an input crop of size $H \times W$ and patch size P, the token count is $N = HW/P^2$. The PCMR gate has shape $N \times D$ and is produced by an MLP whose weights are shared across tokens. Thus the architecture exposes ND routing coordinates while using shared parameters; this is a tensor shape and parameter-sharing property, not a statistical capacity or VC-dimension claim.

H.2 GRADIENT IDENTITIES

For one block, differentiating Eq. (2) with respect to an individual gate entry gives

$$\frac{\partial \mathcal{L}}{\partial G^{(\ell)}_{n,d}} = \frac{\partial \mathcal{L}}{\partial F^{(\ell)}_{n,d}} (\hat{T}^{(\ell)}_{A,n,d} - \hat{T}^{(\ell)}_{B,n,d}). \tag{15}$$

This identity shows that routing updates are driven by both downstream error and cross-modal discrepancy. It does not imply a non-vanishing gradient, because the discrepancy, sigmoid derivative, and subsequent Jacobians can all reduce the gradient.

At the input of the compositor, $\mathbf{C}$, $\mathbf{M}_F$, $\mathbf{R}$, $\mathbf{M}_A$, and $\mathbf{M}_B$ are distinct intermediate tensors. Holding the other compositor inputs fixed, the explicit partial derivative is therefore

$$\left.\frac{\partial \widetilde{\mathbf{M}}}{\partial \mathbf{M}_A}\right|_{\mathbf{C},\mathbf{M}_F,\mathbf{R},\mathbf{M}_B} = \tfrac{1}{2}\mathbf{I}, \tag{16}$$

with the analogous result for $\mathbf{M}_B$. For any shared upstream variable $\mathbf{u}$, however, the complete differential contains all chain-rule paths:

$$\begin{aligned} \mathrm{d}_{\mathbf{u}}\widetilde{\mathbf{M}} = \mathbf{C} \odot \mathrm{d}_{\mathbf{u}}\mathbf{M}_F + (1-\mathbf{C}) \odot \mathrm{d}_{\mathbf{u}}\mathbf{R} \\ + (\mathbf{M}_F - \mathbf{R}) \odot \mathrm{d}_{\mathbf{u}}\mathbf{C} + \tfrac{1}{2}(\mathrm{d}_{\mathbf{u}}\mathbf{M}_A + \mathrm{d}_{\mathbf{u}}\mathbf{M}_B)\,. \end{aligned} \tag{17}$$

The fixed average term thus contributes a direct computational path, but it does not provide a non-zero lower bound on the total gradient because the remaining paths can vanish or cancel it.

H.3 CONDITIONAL PROPERTIES

For fixed gate values, Eq. (2) is affine in the two cross-attended token tensors. This conditional statement does not make the full network or its training objective convex: the gate, attention, FFN, confidence head, decoder, and SSIM terms remain jointly non-convex.

H.4 CWC INTERPRETATION

If the fixed source anchor is omitted, the first two terms of Eq. (6) form a learned channel–spatial interpolation between the routed map and the residual map. This deterministic mixing form does not make $\mathbf{C}$ a calibrated probability, uncertainty estimate, or posterior precision, nor does it guarantee that optimization collapse cannot occur.

I LIMITATIONS

Limitations of the present work include: (1) the fixed patch size $P{=}8$ is a single design point and may be suboptimal for multi-scale scenes; (2) the model assumes pre-registered input pairs; (3) all reported results correspond to a single training run; multi-seed re-training is deferred to future work and is particularly relevant for small test sets such as TNO ($N = 25$); (4) the channel-averaged view of $\mathbf{G}^{(\ell)}$ used to describe routing in Appendix G summarises spatial behaviour but does not directly characterise per-channel diversity, which would require a dedicated interpretability study; (5) training data is restricted to LLVIP plus synthetic multi-focus pairs, so generalisation to the remote-sensing, medical and multi-exposure benchmarks reflects transfer rather than in-domain training, and is correspondingly more conservative in scope.

J ETHICAL CONSIDERATIONS

Infrared-visible image fusion has legitimate applications in autonomous driving, search-and-rescue, and medical imaging. The same technology also enhances surveillance capabilities in civil and military contexts. Improvements in low-light pedestrian detection fidelity—a direct application of IVIF methods including GaIF—can raise privacy and civil-liberties concerns when deployed without appropriate oversight. We encourage responsible deployment in contexts with regulatory accountability and informed consent frameworks.