Beyond Spatial-Domain Supervision: A Relation Constrained Space for Multi-Modal Image Fusion

Zeyu Wang
College of Computer Science and Engineering
Dalian Minzu University
20231578@dlnu.edu.cn

Mingyu Ge
College of Computer Science and Engineering
Dalian Minzu University
2019082204@stu.dlnu.edu.cn

Haiyu Song*
College of Computer Science and Engineering
Dalian Minzu University
shy@dlnu.edu.cn

Haoran Duan*
Department of Automation
Tsinghua University
haoran.duan@ieee.org

Abstract

Multi-modal image fusion (MMIF) aims to form a single image by integrating shared information, preserving complementary cues, and coordinating cross-modal conflicts across modalities. However, due to the absence of ground-truth fused images, existing MMIF supervision commonly uses spatial-domain sources or gradient variants as surrogate ground truths, making the supervision mechanism inherently misaligned with the goal of MMIF and causing pixel-level compromise or modality bias. To address this, we propose a relation-constrained supervision paradigm that moves fusion supervision from the spatial domain to a learned relation space. Instead of forcing the fused image to approximate the sources, we use frozen pretrained representation models as information providers and design a learnable feature adapter to align heterogeneous DINO and CLIP features into a unified supervision space. The adapter infers three relation parameters, namely sharedness, dominance, and coordination radius, which define three losses corresponding to the MMIF's goal. To make this space reliable, we devise a self-supervised contrastive ranking objective tailored to the adapter and couple it with the fusion network through alternating optimization. Extensive experiments show that the proposed supervision space yields significant gains regardless of which mainstream backbone the fusion network adopts, offering a supervision paradigm better aligned with the goal of MMIF. Code: github.com/GMY628/RCS-Fusion.

1 Introduction

Multi-modal Image Fusion (MMIF) combines complementary information from different modalities into one image [30, 31, 46]. Representative tasks include infrared visible fusion [18, 45], which integrates thermal cues with visible textures, and medical image fusion, which combines anatomical and functional information. MMIF also benefits the object detection [19] and medical segmentation [36].

Recent MMIF methods have advanced architectures, cross-modal interaction, learning strategies, and loss design [35, 14, 17, 41]. Despite this, due to the absence of ground-truth (GT) fused images, existing methods typically treat spatial-domain source images or their gradient variants as surrogate targets for loss computation. **This practice has become standard, but introduces a fundamental problem: the supervision mechanism is inherently misaligned with the goal of MMIF**.

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

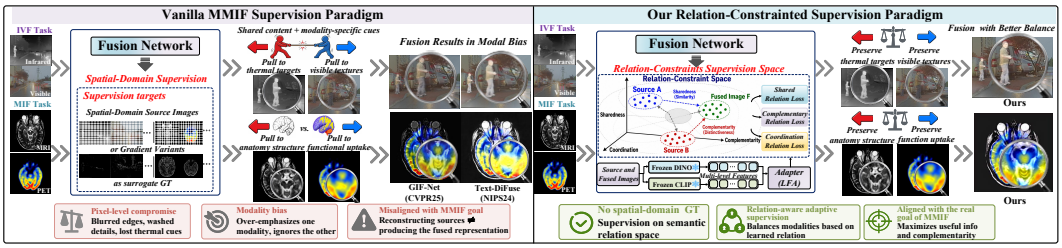


Figure 1: Vanilla MMIF Supervision Paradigm *vs* Our Paradigm. Our method constructs a relation-constraint supervision space tailored for MMIF, leading to better modality balance and fusion quality.

The goal of MMIF is to integrate shared information, preserve modality specific complementary cues, and coordinate conflicts between heterogeneous observations, rather than reconstruct any individual source image. Source images should therefore act as information providers instead of reconstruction targets. Once they are treated as simultaneous GT, the model is forced to satisfy competing constraints from different modalities. This leads to either a pixel-level compromise or a biased result dominated by one modality, as shown in Fig. 1. More importantly, complementary cues, such as infrared thermal targets and visible fine textures, are converted into conflicting supervisory signals.

A natural solution is to move supervision beyond raw spatial domain source images. Recent pretrained representation models [24, 26] provide rich visual features that encode structural and semantic cues. Nevertheless, directly using these features does not solve the problem. Features from different pretrained models are heterogeneous and cannot naturally form a unified supervision space. More critically, ideal MMIF supervision requires not only strong representation capacity, but also explicit relation modeling between modalities: what information is shared, what information is complementary, and how conflicting cues should be coordinated. Without such relation modeling, pretrained features may enrich the representation, but supervision remains insufficiently aligned with fusion.

Motivated by this observation, we propose a relation-constrained supervision paradigm for MMIF. Instead of using source images as crude surrogate ground truths, we construct an adaptive supervision space from frozen pretrained representation models and constrain it according to cross-modal relations. Our core idea is that a proper fusion supervision space should be defined by three fundamental relations: sharedness, complementarity, and coordination (details analysis see Appendix C). Sharedness retains information consistently supported by multiple sources. Complementarity preserves modality specific cues that are informative but weak or absent in the other modality. Coordination prevents conflicting responses from driving the fused result excessively toward either source. Together, these relations provide a supervision principle that matches the intrinsic objective of MMIF.

To implement this principle, we design a learnable feature adapter (LFA) that bridges heterogeneous pretrained features and the proposed supervision space. LFA consists of a Feature Alignment Module and a Relation Reasoning Module. The former projects intermediate features from frozen DINO and CLIP into a unified feature space. Within this space, we tailor three losses, namely sharedness, complementarity, and coordination losses, to supervise fusion. The latter predicts three relation parameters that adaptively weight these losses, allowing supervision to reflect the cross-source relationship of each image pair rather than fixed manual coefficients.

Besides, training the adapter is critical yet challenging. An unreliable adapter would produce misleading supervision and thus undermine any fusion model trained with it. However, no ready made training set or established strategy exists for learning such a relation-aware adapter, creating a circular dependency between adapter learning and fusion supervision. We therefore tailor a self-supervised contrastive training scheme to the adapter and couple it with alternating optimization. Specifically, we pretrain the adapter with a ranking loss on relation scores computed from its outputs, so that matched pairs (positive pairs) score higher than unmatched ones (negative pairs). This enables LFA to learn cross-modal relation estimation without explicit relation annotations. Before alternating optimization, the fusion network is first trained with a conventional fusion loss to reach a baseline level of fusion performance. We then alternate between freezing the adapter to supervise the fusion network with relation losses and freezing the fusion network to train the adapter with fused images under the same ranking objective, encouraging each fused image to score higher with its matched sources than with mismatched ones. This strategy enables effective training of both the adapter and

the fusion network. Our contributions are summarized as follows:
•We revisit MMIF supervision from the objective of fusion and propose a relation-constrained paradigm beyond spatial domain source images as surrogate ground truths.
•We define the supervision space through sharedness, complementarity, and coordination, aligning supervision with the requirements of fusion.
•We design a learnable feature adapter that aligns heterogeneous pretrained features and predicts relation parameters to automatically construct the supervision space.
•We propose a self-supervised relation ranking strategy and an alternating optimization scheme to jointly train the adapter and fusion network, forming a bidirectional optimization mechanism.

2 Related Work

Supervision Paradigms for MMIF. Recent MMIF research has progressed from reconstruction-based frameworks to recent Transformer, diffusion, and language-guided models [53, 13, 52]. Because no physical ground-truth fused image exists, their performance remains highly dependent on supervision loss design [2, 1]. Existing methods mainly follow source-preservation, perceptual or adversarial, task-driven, and semantic or language-guided paradigms [19, 37, 45]. However, they still use spatial-domain source images as GT or proxy targets for loss computation. This misaligns optimization with the objective of MMIF and often causes pixel-level compromise or modality bias. In contrast, our method abandons this long-standing practice and uses powerful frozen pretrained representation models to build a supervision paradigm better aligned with the goal of MMIF.

Pretrained Visual Knowledge for MMIF. Recent advances in pretrained representation models have made large-scale visual knowledge a promising resource for image fusion. Models such as CLIP and DINOv2 learn rich structural and semantic cues from massive data, and their representations have shown strong transferability in low-level vision tasks [39, 7, 11, 12]. In image fusion, recent studies have explored knowledge-aware or representation-driven fusion through semantic guidance [47, 1, 10], text descriptions [42, 5], or self-supervised representation learning [51, 50]. However, these methods mainly use pretrained knowledge as auxiliary features or side information and do not turn it into direct supervision for fusion. In contrast, our method abandons this long-standing practice and uses external knowledge from powerful frozen pretrained representation models to build a supervision paradigm better aligned with the fundamental objective of MMIF.

3 Methodology

Overview. As shown in Fig. 2, our method consists of three tightly coupled components: a learnable feature adapter (LFA), relation-constrained supervision (RCS) for MMIF, and a tailored training strategy. The LFA maps heterogeneous multi-level features from frozen DINO [24] and CLIP [26] into a unified supervision space and infers source relations. Based on these outputs, we construct the RCS with three losses for shared information integration, complementary cues preservation, and conflict coordination. Finally, we train the LFA with a self-supervised contrastive learning strategy and optimize it alternately with the fusion net. Appendix B gives the formal problem statement.

3.1 Learnable Feature Adapter (LFA)

The learnable feature adapter (LFA) contains two components: a Feature Alignment Module (FAM) and a Relation Reasoning Module (RRM). For each source image $X \in \{A, B\}$, frozen DINO and CLIP extract three feature maps from DINO layer 4, CLIP layer 6, and DINO layer 8, denoted by $f_D^4(X)$, $f_C^6(X)$, and $f_D^8(X)$, respectively. FAM takes these three feature maps as input and outputs an aligned representation y_X in the unified supervision space. RRM then takes the aligned source pair (y_A, y_B) as input and outputs two pair-conditioned source representations, z_A and z_B, together with three relation parameters: sharedness $m_{A,B}$, dominance $d_{A,B}$, and coordination radius $r_{A,B}$.

Feature Alignment Module. To address the mismatch in spatial layout and representation structure across models and feature depths, FAM processes the multi-level features of a source pair (A, B) with a shared projection-and-alignment pipeline. Specifically, for each source image $X \in \{A, B\}$, FAM projects the three feature maps to the same channel dimension and a common token grid:

$$\tilde{f}_D^4(X) = P_D^4\big(f_D^4(X)\big), \quad \tilde{f}_C^6(X) = P_C^6\big(f_C^6(X)\big), \quad \tilde{f}_D^8(X) = P_D^8\big(f_D^8(X)\big), \tag{1}$$

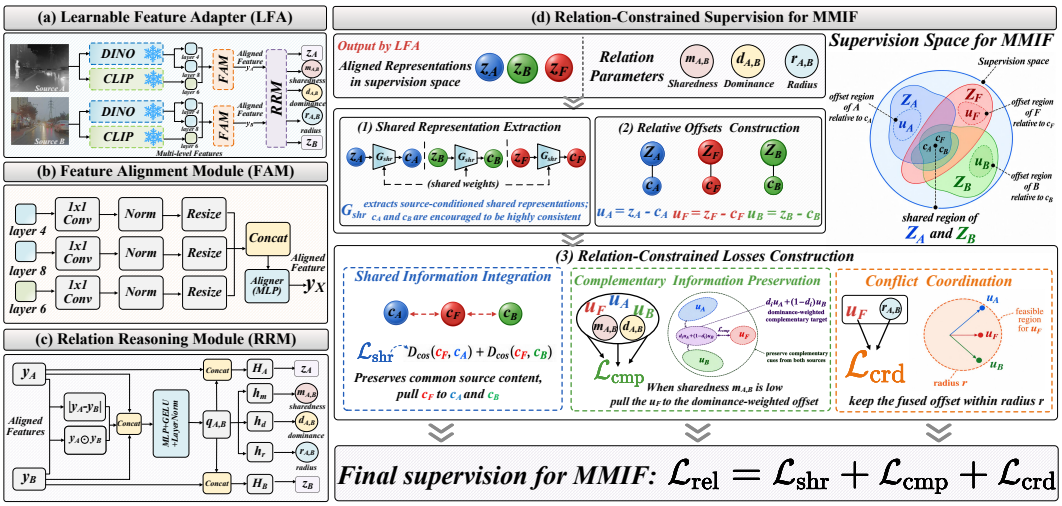


Figure 2: Overview of the Learnable Feature Adapter and Relation-Constrained Supervision.

where P_D^4, P_C^6, and P_D^8 denote the corresponding projection operators, and $\tilde{f}_D^4(X)$, $\tilde{f}_C^6(X)$, and $\tilde{f}_D^8(X)$ denote the projected features. Each projector is implemented as a 1×1 convolution followed by normalization and bilinear resizing, so that the three inputs share the same token layout. The projected features are then flattened into token sequences, concatenated along the channel dimension, and fused by a cross-model aligner:

$$y_X = G_{\text{align}}\Big([\tilde{f}_D^4(X), \tilde{f}_C^6(X), \tilde{f}_D^8(X)]\Big), \quad X \in \{A, B\}, \tag{2}$$

where $[\cdot]$ denotes concatenation and G_{align} denotes a lightweight MLP-based aligner. Applying this shared pipeline to the two source images produces the aligned representations y_A and y_B.

Relation Reasoning Module. Given the aligned source representations y_A and y_B produced by FAM, RRM infers the relation of the source pair in the unified supervision space. To support this relation inference, RRM forms a pair input by concatenating the two aligned source representations, their element-wise absolute difference, and their element-wise product. The absolute difference captures the magnitude of source discrepancy, while the element-wise product reflects their feature-wise agreement. This pair input is first encoded into a shared pair feature $q_{A,B}$ and then projected to two pair-conditioned source representations z_A and z_B:

$$q_{A,B} = E_{\text{pair}}([y_A, y_B, |y_A - y_B|, y_A \odot y_B]), \;\; z_A = H_A([q_{A,B}, y_A]), \;\; z_B = H_B([q_{A,B}, y_B]), \tag{3}$$

where E_{pair} denotes the pair encoder, and H_A and H_B denote two source-specific projection heads. In our implementation, E_{pair} is a lightweight MLP with GELU and LayerNorm, and H_A and H_B are two linear heads followed by normalization.

Based on the shared pair feature $q_{A,B}$, RRM further predicts three relation parameters:

$$m_{A,B} = \text{Sigmoid}(h_m(q_{A,B})), \quad d_{A,B} = \text{Sigmoid}(h_d(q_{A,B})), \quad r_{A,B} = \text{Softplus}(h_r(q_{A,B})), \tag{4}$$

where h_m, h_d, and h_r are lightweight prediction heads, each implemented by a linear layer, GELU, and LayerNorm. Accordingly, $m_{A,B}$ measures sharedness, $d_{A,B}$ quantifies the tendency toward source A in the non-shared part, with $1 - d_{A,B}$ giving the complementary contribution of source B, and $r_{A,B}$ defines the coordination radius. Together, the representations z_A and z_B and the relation parameters $m_{A,B}$, $d_{A,B}$, and $r_{A,B}$ are used to construct the subsequent supervision space for MMIF.

3.2 Relation-Constrained Supervision (RCS) for MMIF

We now use the outputs of LFA to construct relation-constrained supervision for MMIF in the unified supervision space. For a source pair (A, B), RRM provides two pair-conditioned source representations, z_A and z_B, together with three relation parameters, $m_{A,B}$, $d_{A,B}$, and $r_{A,B}$. We obtain z_F by encoding the self-pair (F, F) with the same pair-conditioned LFA.

Since the source images contain both shared and non-shared information, a fixed target representation is insufficient for fusion supervision. We introduce a pretrained shared token encoder G_{shr}, which is applied with shared weights to the aligned source representations to extract their shared components:

$$c_A = G_{\mathrm{shr}}(z_A), \qquad c_B = G_{\mathrm{shr}}(z_B). \tag{5}$$

where c_A and c_B denote source-conditioned shared representations. During pretraining, G_{shr} is encouraged to make them highly consistent, so they retain information commonly supported by both modalities and jointly define the consensus reference. The details are in the Appendix K.

Next, we define the relative offsets of the two source and fused representations as:

$$u_A = z_A - c_A, \quad u_B = z_B - c_B, \quad u_F = z_F - c_F, \tag{6}$$

where $c_F = G_{\mathrm{shr}}(z_F)$ denotes the shared representation extracted from the fused representation. u_A and u_B denote the source offsets, and u_F denotes the fusion offset. These quantities allow us to define three losses for shared information integration, complementary information preservation, and conflict coordination.

Shared Information Integration. The first requirement of MMIF is to preserve information shared by the two source images. Thus, the fused shared representation should stay close to both shared components rather than drift toward either modality. Accordingly, the shared loss takes the form:

$$\mathcal{L}_{\mathrm{shr}} = \frac{1}{N}\sum_{i=1}^{N}\left[\mathcal{D}_{\cos}(c_{F,i}, c_{A,i}) + \mathcal{D}_{\cos}(c_{F,i}, c_{B,i})\right]. \tag{7}$$

where N denotes the number of tokens, i indexes the token position, and $\mathcal{D}_{\cos}(\cdot,\cdot)$ denotes the cosine distance. Reducing the loss requires $c_{F,i}$ to align with the shared representations $c_{A,i}$ and $c_{B,i}$, which preserves common source content without introducing modality bias.

Complementary Information Preservation. Fusion should not collapse all information to the shared cues. In regions with low sharedness, the fused result should preserve useful non-shared cues from the two sources. The complementary loss is given by:

$$\mathcal{L}_{\mathrm{cmp}} = \frac{1}{N}\sum_{i=1}^{N}(1-m_i)\;\mathcal{D}_{\cos}(u_{F,i},\, d_i u_{A,i} + (1-d_i)\, u_{B,i}), \tag{8}$$

where $d_i u_{A,i} + (1-d_i)\, u_{B,i}$ denotes the dominance-weighted complementary target offset. The factor $1-m_i$ makes this term focus on non-shared regions. When the loss decreases, the fused offset is encouraged to follow the dominance-aware direction defined by the two source offsets, which helps preserve complementary information from both modalities.

Conflict Coordination. Preserving complementary information alone is insufficient under strong cross-modal conflict, because the fused representation may still be driven excessively toward one source. To prevent this, we introduce a coordination constraint that keeps the fused offset within an adaptive range. The coordination loss is defined as:

$$\mathcal{L}_{\mathrm{crd}} = \frac{1}{N}\sum_{i=1}^{N}\max(0, \|u_{F,i}\|_2 - r_i), \tag{9}$$

where r_i denotes the coordination radius at token i. This hinge form imposes no penalty when the fused offset stays within the allowed radius. It only penalizes excessive deviation, thereby limiting conflict-driven bias while preserving flexibility for useful complementary cues.

The final relation-constrained supervision is defined as:

$$\mathcal{L}_{\mathrm{rel}} = \mathcal{L}_{\mathrm{shr}} + \mathcal{L}_{\mathrm{cmp}} + \mathcal{L}_{\mathrm{crd}}. \tag{10}$$

In this way, fusion supervision is no longer defined by direct approximation to the source images in the spatial domain. Instead, it is defined by whether the fused representation satisfies the shared, complementary, and coordination requirements induced by source relations in the proposed supervision space. Appendix D analyzes the effects of relation parameters on the supervision losses.

3.3 Training Strategy

Self-Supervised Contrastive Learning for the Adapter. Before using LFA to supervise the fusion network, we pre-train the adapter with a self-supervised contrastive ranking objective to make the

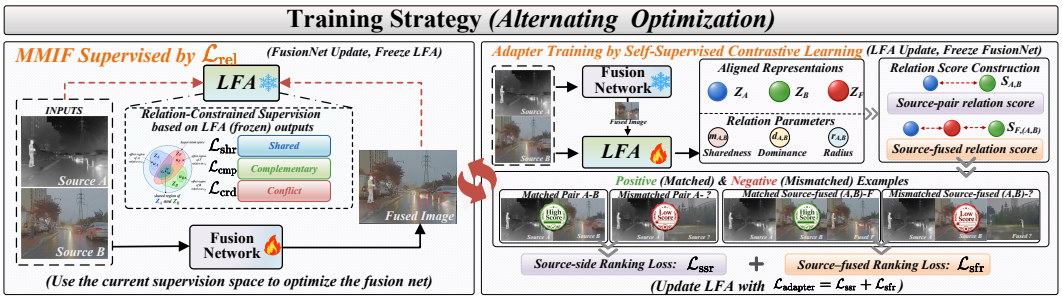


Figure 3: Overview of the Alternating Optimization Strategy.

learned supervision space discriminative for source relations. Since LFA outputs both pair-conditioned representations and relation parameters, the score used for training should depend on both parts. For a source pair (A, B), we define the source-pair relatedness score as:

$$s_{A,B} = \frac{\sum_{i=1}^{N} w_i \cos(z_{A,i}, z_{B,i})}{\sum_{i=1}^{N} w_i + \epsilon}, \quad w_i = \sigma(\lambda_m m_i + \lambda_d |2d_i - 1| - \lambda_r r_i), \tag{11}$$

where N denotes the number of tokens, ϵ is a small constant, $\cos(\cdot, \cdot)$ denotes cosine similarity, and λ_m, λ_d, and λ_r are weighting coefficients. In this score, $\cos(z_{A,i}, z_{B,i})$ measures the local agreement between the two pair-conditioned representations, while w_i determines how much token i contributes to the final score. The design of w_i follows the roles of the three relation parameters: a larger m_i gives higher weight to shared regions, a larger $|2d_i - 1|$ gives higher weight to regions with clearer dominance, and a larger r_i reduces the weight of regions that require a wider coordination range. As a result, $s_{A,B}$ becomes high only when z_A and z_B are locally consistent at tokens where m_i, d_i, and r_i indicate reliable source relations. For negative source pairs such as (A, B^-) and (A^-, B), the score is computed in the same way after replacing the corresponding source. Appendix E further explains the relatedness scores and ranking objective.

We first use source-side ranking to make the relation space discriminative at the source-pair level. Let (A, B^-) and (A^-, B) denote two negative pairs formed by replacing one source with an unpaired sample. The source-side ranking loss is:

$$\mathcal{L}_{\text{ssr}} = \frac{1}{2}\Big(\max(0, \delta_{\text{src}} - s_{A,B} + s_{A,B^-}) + \max(0, \delta_{\text{src}} - s_{A,B} + s_{A^-,B}) \Big), \tag{12}$$

where δ_{src} denotes the source-side margin. This margin requires the matched pair to score higher than each mismatched pair by a non-trivial gap, instead of only being slightly larger.

Source-side ranking alone is insufficient, because LFA is ultimately used to supervise fused images. Without source–fused ranking, the learned relation space may not remain reliable when fused images are introduced. We therefore define the relatedness between the fused image and its source pair as:

$$s_{F,(A,B)} = \frac{\sum_{i=1}^{N} w_i \Big(m_i c_{shr} + (1 - m_i) \cos\big(u_{F,i},\, d_i u_{A,i} + (1 - d_i) u_{B,i}\big) \Big)}{\sum_{i=1}^{N} w_i + \epsilon}, \tag{13}$$

where $c_{shr} = \cos(c_{F,i}, c_{A,i}) + \cos(c_{F,i}, c_{B,i})$, $u_{A,i}$, $u_{B,i}$, and $u_{F,i}$ are defined in the previous subsection. The first term checks whether the fused representation is close to the shared information, while the second term checks whether it follows the dominance-aware complementary offset. Based on this score, the source–fused ranking loss is:

$$\mathcal{L}_{\text{sfr}} = \frac{1}{2}\Big(\max(0, \delta_{\text{fuse}} - s_{F,(A,B)} + s_{F,(A,B^-)}) + \max(0, \delta_{\text{fuse}} - s_{F,(A,B)} + s_{F,(A^-,B)}) \Big), \tag{14}$$

where δ_{fuse} denotes the fused-side margin, analogous to δ_{src} in source-side ranking. The overall adapter objective is:

$$\mathcal{L}_{\text{adapter}} = \mathcal{L}_{\text{ssr}} + \mathcal{L}_{\text{sfr}}. \tag{15}$$

Alternating optimization. Before alternating optimization, the fusion network is first warmed up with a conventional spatial-domain fusion loss to obtain a reasonable baseline performance. Specifically, we adopt the commonly used intensity and gradient constraints:

$$\mathcal{L}_{\text{warm}} = \alpha_1 \|F - \max(A, B)\|_1 + \alpha_2 \|\nabla F - \max(\nabla A, \nabla B)\|_1, \tag{16}$$

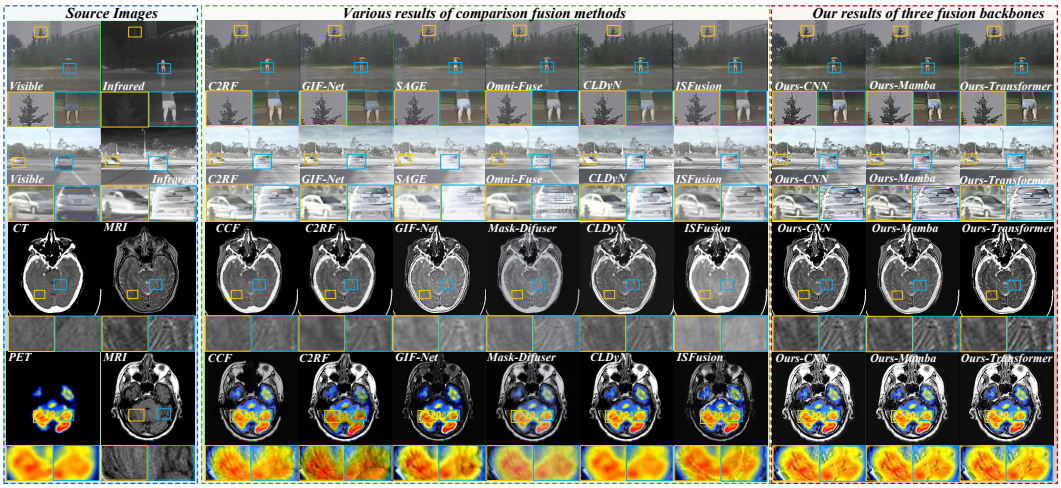


Figure 4: Qualitative comparison for IVF (top two rows) and MIF (bottom two rows).

Table 1: Quantitative comparison for IVF and MIF task. The best, second best, and third best are highlighted in **red**, **blue**, and **green**, respectively. "Ours(CNN/Mamba/Transformer)" denotes CNN, Mamba, and Transformer fusion backbones trained under our relation-constrained supervision.

IVF Task		RoadScene							M^{3}FD							MSRS						
Methods	Pub./Year	$AG\uparrow$	$EN\uparrow$	$SD\uparrow$	$MI\uparrow$	$Q_{AB/F}\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$	$AG\uparrow$	$EN\uparrow$	$SD\uparrow$	$MI\uparrow$	$Q_{AB/F}\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$	$AG\uparrow$	$EN\uparrow$	$SD\uparrow$	$MI\uparrow$	$Q_{AB/F}\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$
EMMA	CVPR 24	4.779	49.127	46.777	2.809	0.436	0.432	0.355	5.871	57.569	42.377	3.771	0.589	0.512	0.460	3.788	40.539	40.590	4.164	0.541	0.760	0.496
Text-DiFuse	NIPS 24	3.962	42.863	48.966	1.961	0.140	0.288	0.052	1.861	20.580	40.568	2.054	0.066	0.240	0.332	2.792	30.723	39.679	1.310	0.118	0.288	0.038
Tc-MoA	CVPR 24	4.192	44.883	42.610	1.043	0.413	0.556	0.403	5.180	54.156	39.030	3.333	0.651	0.891	0.513	3.618	38.414	42.335	3.493	0.630	1.090	0.483
ReFusion	IJCV 24	5.357	59.336	48.229	2.726	0.437	0.566	0.397	5.338	57.331	40.803	3.548	0.637	0.731	0.489	3.813	40.641	42.890	4.102	0.442	1.085	0.576
C2RF	IJCV 25	4.989	52.270	45.576	2.720	0.519	0.565	0.388	4.779	50.329	39.348	2.498	0.223	0.287	0.399	3.474	37.153	41.348	1.729	0.185	0.330	0.092
GIF-Net	CVPR 25	5.705	59.855	45.000	2.236	0.352	0.365	0.252	5.435	52.096	42.714	2.529	0.471	0.319	0.319	3.379	36.267	32.910	1.926	0.367	0.396	0.277
SAGE	CVPR 25	3.715	39.057	46.384	1.035	0.362	0.451	0.332	4.885	50.959	40.930	2.851	0.592	0.599	0.434	3.342	35.157	37.485	2.990	0.539	0.714	0.435
Omni-Fuse	TPAMI 25	3.206	35.212	44.830	2.187	0.303	0.396	0.267	4.514	49.067	41.029	3.208	0.430	0.362	0.277	3.091	34.123	40.043	2.664	0.361	0.412	0.217
CLDyN	CVPR 26	4.918	51.612	40.644	1.827	0.417	0.475	0.329	5.824	57.312	35.254	3.235	0.689	0.756	0.520	3.568	37.882	39.377	2.111	0.661	0.867	0.484
ISFusion	CVPR 26	5.320	54.393	35.946	2.345	0.522	0.484	0.306	5.080	52.927	36.718	2.188	0.511	0.422	0.343	4.124	42.850	34.100	3.303	0.491	0.550	0.357
Ours(CNN)	- -	5.701	59.876	49.673	2.962	0.541	0.604	0.411	5.478	57.921	42.315	4.238	0.664	1.041	0.529	3.842	40.982	43.517	4.301	0.651	1.337	0.509
Ours(Mamba)	- -	5.858	60.742	50.214	3.084	0.552	0.621	0.417	5.602	58.736	42.982	4.612	0.671	1.102	0.533	3.902	41.427	43.918	4.493	0.664	1.462	0.518
Ours(Transformer)	- -	5.861	61.069	50.581	3.179	0.564	0.643	0.425	5.666	59.158	43.407	4.837	0.679	1.157	0.540	3.953	41.854	44.231	4.651	0.678	1.576	0.526
MIF Task		CT-MRI							MRI-PET							MRI-SPECT						
Methods	Pub./Year	$AG\uparrow$	$EN\uparrow$	$SD\uparrow$	$MI\uparrow$	$Q_{AB/F}\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$	$AG\uparrow$	$EN\uparrow$	$SD\uparrow$	$MI\uparrow$	$Q_{AB/F}\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$	$AG\uparrow$	$EN\uparrow$	$SD\uparrow$	$MI\uparrow$	$Q_{AB/F}\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$
EMMA	CVPR 24	7.317	76.268	72.599	3.279	0.551	0.153	0.316	7.116	75.114	76.489	2.448	0.541	0.143	0.318	4.689	48.655	62.720	2.361	0.516	0.223	0.337
Text-DiFuse	NIPS 24	6.831	70.390	66.701	3.159	0.499	0.155	0.264	8.581	88.178	66.353	2.331	0.489	0.144	0.312	5.935	60.168	58.726	2.225	0.539	0.226	0.365
Tc-MoA	CVPR 24	6.888	72.785	72.894	3.317	0.592	0.162	0.369	7.148	75.676	69.960	2.450	0.640	0.195	0.469	4.251	44.348	55.776	2.333	0.426	0.272	0.356
ReFusion	IJCV 24	8.731	88.753	76.113	3.176	0.538	0.223	0.356	7.880	80.686	72.422	2.486	0.539	0.212	0.458	5.490	55.098	61.274	2.310	0.580	0.557	0.372
C2RF	IJCV 25	7.978	81.643	76.192	2.953	0.332	0.171	0.131	7.933	81.310	70.626	2.183	0.458	0.216	0.284	5.369	54.366	59.497	2.257	0.547	0.400	0.309
GIF-Net	CVPR 25	9.544	94.839	70.216	3.067	0.466	0.128	0.236	6.489	67.343	61.177	2.299	0.348	0.105	0.245	4.742	48.710	51.468	2.244	0.432	0.176	0.304
CCF	NIPS 24	6.724	67.686	70.568	3.284	0.546	0.163	0.283	6.461	65.927	65.437	2.493	0.512	0.193	0.273	4.398	44.729	64.837	2.241	0.526	0.285	0.391
Mask-Difuser	TPAMI 25	5.984	60.253	69.600	3.299	0.484	0.166	0.213	6.126	61.783	62.930	2.706	0.572	0.200	0.440	4.164	41.390	58.787	2.127	0.548	0.303	0.302
CLDyN	CVPR 26	7.777	80.208	73.716	3.245	0.597	0.206	0.394	7.618	79.791	72.779	2.466	0.670	0.220	0.490	4.919	50.583	63.594	2.870	0.528	0.330	0.609
ISFusion	CVPR 26	8.944	88.753	72.945	2.953	0.556	0.171	0.315	8.322	81.579	66.199	2.171	0.579	0.177	0.349	5.531	54.280	58.691	2.169	0.570	0.267	0.393
Ours(CNN)	- -	9.558	94.876	75.642	3.352	0.611	0.249	0.397	9.127	96.512	76.938	2.514	0.604	0.219	0.463	6.145	63.876	66.873	2.344	0.581	0.423	0.396
Ours(Mamba)	- -	9.573	95.214	76.308	3.421	0.626	0.267	0.443	9.412	97.864	77.614	2.592	0.619	0.228	0.474	6.382	65.214	67.942	2.401	0.594	0.451	0.405
Ours(Transformer)	- -	9.608	95.842	76.527	3.548	0.637	0.290	0.411	9.690	99.031	78.072	2.667	0.627	0.235	0.481	6.571	66.683	68.496	2.448	0.608	0.464	0.413

where ∇ denotes the Sobel gradient operator and $\|\cdot\|_1$ denotes the L_1 norm, while α_1 and α_2 are hyperparameters. This loss is used only for fusion-network warm-up training and is removed once alternating optimization begins.

Next, we switch from conventional fusion loss to the proposed relation-constrained training scheme. The adapter constructs the supervision space, while the fusion network produces the fused images on which this supervision is applied. If either side is kept fixed throughout training, the learned supervision space and the fusion results may remain mismatched. We therefore optimize them alternately. During the adapter-update stage, the fusion network is frozen, and the current fused image F is used to optimize LFA with $\mathcal{L}_{\text{adapter}}$. During the fusion-update stage, LFA is frozen, and the fusion network is optimized with $\mathcal{L}_{\text{rel}}$. This alternating process allows the adapter to refine the supervision space with current fusion results, while the fusion network progressively adapts to the updated supervision. Appendix A provides the full training procedure.

4 Experiments

Datasets and Backbones. We evaluate our method on IVF and MIF tasks. For IVF, we train on 1,000 pairs MSRS [21] and test on held-out MSRS (100 pairs), M^{3}FD [16] (100 pairs), and RoadScene [38] (50 pairs). For MIF, we use the Harvard Medical Dataset [45] for CT-MRI, MRI-PET, and MRI-SPECT fusion. **Since our supervision paradigm is architecture-independent, we**

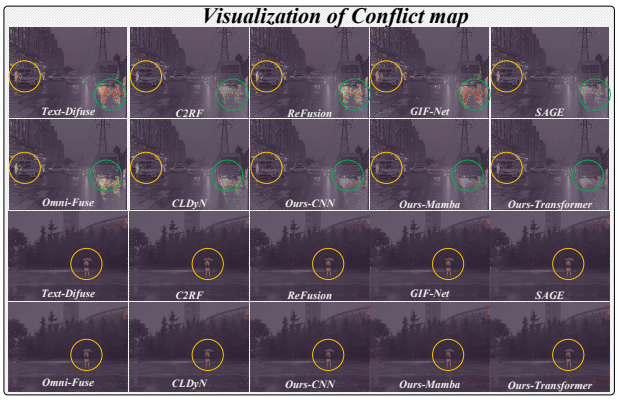


Figure 5: Visualization of conflict map. Warmer colors denote higher source preservation update conflicts, while cooler denote lower conflicts.

Table 2: Quantitative results on source preservation update conflicts in high-disagreement regions on the M^3FD dataset. CR, CS, and UI denote conflict ratio, conflict strength, and update imbalance, respectively.

Method	CR (%) ↓	CS ↓	UI ↓
Text-DiFuse	27.25	0.806	0.4527
C2RF	16.04	0.715	0.3379
ReFusion	20.91	0.784	0.3865
GIF-Net	18.36	0.692	0.3514
SAGE	15.28	0.641	0.3196
Omni-Fuse	17.63	0.667	0.3421
CLDyN	13.72	0.584	0.2968
Ours (CNN)	**7.42**	**0.315**	**0.218**
Ours (Mamba)	**6.37**	**0.287**	**0.198**
Ours (Transformer)	**4.77**	**0.265**	**0.176**

adopt CNN- [48], Mamba- [20], and Transformer-based [44] fusion backbones and provide their details in the Appendix K.

Training Details. For IVF, we train on MSRS with 1000 pairs. For MIF, we use 200 training pairs on Harvard Medical Dataset. For each task, the fusion net is first warmed up for 20 epochs with the conventional fusion loss. The adapter is then pretrained for 50 epochs with the self-supervised ranking objective. After initialization, alternating optimization is performed for 5 rounds. In each round, the adapter is updated for 30 epochs with the fusion network frozen, and the fusion network is updated for 20 epochs with the adapter frozen. During the fusion-update stage, only the proposed relation-constrained supervision is used. All these training stages use AdamW with an initial learning rate of 3×10^{-4}, batch size 4, and no weight decay. The learning rate is halved every 10 epochs. All experiments are conducted on an NVIDIA RTX 4090 GPU. Parameter settings are in the Appendix K.

Comparison methods and metrics. We compare with EMMA [51], Text-DiFuse [45], Tc-MoA [54], ReFusion [2], C2RF [31], GIF-Net [6], SAGE [37], Omni-Fuse [46], CCF [4], Mask-Difuser [30], CLDyN [41], and ISFusion [34], using AG [8], EN [28], SD [27], MI [25], $Q_{AB/F}$ [40], Q_M [33], and Q_P [49] as metrics.

4.1 Comparison with SOTA

Quantitative Comparison. Table 1 compares our method with recent SOTA methods on IVF and MIF tasks. "Ours(CNN/Mamba/Transformer)" denotes different fusion backbones trained under our relation-constrained supervision. All three variants achieve competitive or superior results across datasets, with the Transformer variant obtaining the best overall performance. These results indicate that relation-space supervision provides a more task-aligned optimization target than spatial-domain source approximation. **Qualitative Comparison.** Fig. 4 shows that existing methods often suffer from modality bias, either weakening infrared targets or visible textures in IVF, and either suppressing functional uptake or distorting anatomical boundaries in MIF. In contrast, our method preserves salient targets, structural details, and functional regions more faithfully, producing clearer and more balanced fusion results. Full qualitative comparisons are provided in Appendix F.

4.2 Analysis of Source-Preservation Update Conflicts

Following task-gradient conflict analysis in multi-task optimization [43, 15, 23, 29], we view infrared and visible source preservation as two MMIF analysis tasks. This analysis is used only for evaluation. For a fused result F^m, we define $\mathcal{L}_s^{sp} = \|F^m - I_s\|_1 + \lambda_g(\|D_x F^m - D_x I_s\|_1 + \|D_y F^m - D_y I_s\|_1)$, where $s \in \{ir, vis\}$ and D_x, D_y are Sobel operators. The induced update is $u_s^m = -\partial \mathcal{L}_s^{sp} / \partial F^m$, and patch-level agreement is measured by $\rho_p^m = \cos(u_{ir,p}^m, u_{vis,p}^m)$. A negative ρ_p^m indicates opposite local updates from the two sources. Specifically, we evaluate conflicts on the high-disagreement patch set Ω, which is selected only from source-image intensity and gradient discrepancies and is shared by all methods. On Ω, we report the conflict ratio $\mathrm{CR} = |\Omega|^{-1} \sum_{p \in \Omega} \mathbb{I}(\rho_p^m < 0)$ as a percentage,

Table 3: Ablation study of the proposed method on the IVF RoadScene and MIF MRI-PET datasets. The best results are highlighted in **bold**. More ablation variants are provided in the appendix.

Group	ID	Description	IVF on RoadScene							MIF on MRI-PET						
			$AG\uparrow$	$EN\uparrow$	$SD\uparrow$	$MI\uparrow$	$Q_{AB/F}\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$	$AG\uparrow$	$EN\uparrow$	$SD\uparrow$	$MI\uparrow$	$Q_{AB/F}\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$
(a)	(1)	w/o LFA	5.247	58.932	47.286	2.861	0.506	0.589	0.372	8.721	94.863	72.418	2.321	0.552	0.196	0.414
	(2)	w/o FAM	5.516	60.217	49.103	3.041	0.537	0.618	0.401	9.184	97.326	75.861	2.512	0.594	0.218	0.452
	(3)	w/o RRM	5.438	59.846	48.624	2.982	0.529	0.611	0.394	9.062	96.814	75.104	2.463	0.586	0.213	0.444
	(4)	w/o RCS	5.196	58.714	47.635	2.803	0.497	0.581	0.363	8.604	94.217	71.936	2.276	0.541	0.189	0.405
(b)	(1)	w/o $\mathcal{L}_{\text{shr}}$	5.642	60.384	49.427	3.086	0.545	0.626	0.410	9.346	97.942	76.412	2.574	0.607	0.224	0.461
	(2)	w/o $\mathcal{L}_{\text{cmp}}$	5.471	60.092	48.891	2.947	0.523	0.613	0.392	9.118	97.156	75.386	2.438	0.581	0.216	0.441
	(3)	w/o $\mathcal{L}_{\text{crd}}$	5.573	60.263	49.018	3.018	0.532	0.607	0.386	9.241	97.624	75.927	2.496	0.592	0.211	0.433
(c)	(1)	w/o $\mathcal{L}_{\text{ssr}}$	5.324	59.463	48.102	2.902	0.514	0.598	0.379	8.891	95.842	73.864	2.384	0.567	0.204	0.425
	(2)	w/o $\mathcal{L}_{\text{sfr}}$	5.386	59.721	48.406	2.918	0.517	0.602	0.381	8.976	96.214	74.293	2.406	0.573	0.207	0.429
	(3)	fixed adapter	5.452	59.984	48.713	2.971	0.526	0.609	0.389	9.087	96.936	75.021	2.451	0.583	0.212	0.438
Full Model (Transformer)			**5.861**	**61.069**	**50.581**	**3.179**	**0.564**	**0.643**	**0.425**	**9.690**	**99.031**	**78.072**	**2.667**	**0.627**	**0.235**	**0.481**

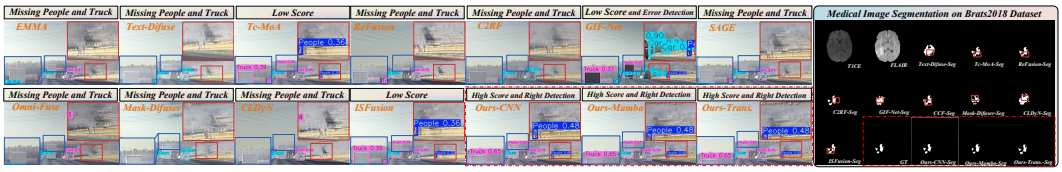


Figure 6: Qualitative results on downstream tasks (Object detection and medical image segmentation).

conflict strength $\text{CS} = |\Omega|^{-1}\sum_{p\in\Omega}[-\rho_p^m]_+$, and update imbalance $\text{UI} = |\Omega|^{-1}\sum_{p\in\Omega}|\|u_{ir,p}^m\|_2 - \|u_{vis,p}^m\|_2|/(\|u_{ir,p}^m\|_2 + \|u_{vis,p}^m\|_2 + \epsilon)$. Lower values indicate fewer conflicts and better source-side update balance. Details of Ω and the metrics see Appendix G. As shown in Fig. 5 and Table 2, existing methods produce stronger high-conflict responses in high-disagreement regions. Ours reduce CR, CS, and UI overall. These results suggest that RCS coordinates cross-modal fusion more stably, and preserves useful source cues with fewer contradictory gradients.

4.3 Ablation Study

We conduct ablation studies on IVF and MIF datasets to verify the effectiveness of our main designs. Key results are shown in Table 3, and more variants are provided in the Appendix I.

(a) Effect of LFA. LFA constructs the supervision space from frozen DINO and CLIP features. We evaluate its role through four variants: (1) removing LFA, (2) removing FAM, (3) removing RRM, and (4) replacing RCS with conventional source preservation. Results in Table 3 (a) show clear performance drops in all cases, especially when LFA or RCS is removed. This shows that pretrained features need task-oriented alignment and relation reasoning for effective fusion supervision.

(b) Effect of RCS. RCS defines fusion supervision through shared integration, complementary preservation, and conflict coordination. We examine its role by removing $\mathcal{L}_{\text{shr}}$, $\mathcal{L}_{\text{cmp}}$, and $\mathcal{L}_{\text{crd}}$, respectively. Results in Table 3 (b) show consistent performance drops in all cases. This shows that adaptive relation constraints are essential for task-aligned fusion supervision.

(c) Effect of Adapter Training. Adapter training makes the learned supervision space discriminative and compatible with fused images. We evaluate its role through three variants: (1) removing $\mathcal{L}_{\text{ssr}}$, (2) removing $\mathcal{L}_{\text{sfr}}$, and (3) fixing the adapter after pretraining. Results in Table 3 (c) show clear degradation across these variants. This confirms that the adapter should be pretrained with relation ranking and alternately refined with the fusion network.

4.4 Downstream Tasks

We further evaluate downstream perception using fused images. For IVF, all methods are tested with the same pretrained YOLOv12 [32] detector without fine-tuning; for MIF, medical fusion results are evaluated with UniverSeg [3]. As shown in Fig. 6 and Tables 4–5, our method achieves more reliable detection and segmentation results than competing methods. These gains indicate that relation-constrained supervision better preserves task-relevant cues and improves downstream utility.

Table 4: Quantitative results of various methods for object detection task on the M^3FD [16].

Object Detection	People	Car	Bus	Motorcycle	Lamp	Truck	mAP
Infrared	0.304	0.422	0.024	0.549	0.000	0.112	0.235
Visible	0.515	0.603	0.691	0.455	0.585	0.491	0.557
EMMA	0.450	0.617	0.457	0.469	0.465	0.517	0.496
Text-DiFuse	0.547	0.614	0.781	0.565	0.632	0.545	0.614
Tc-MoA	0.550	0.687	0.757	0.459	0.565	0.317	0.556
ReFusion	0.489	0.663	0.522	0.538	0.597	0.663	0.579
GIF-Net	**0.695**	0.782	0.756	0.528	0.565	0.477	0.634
SAGE	0.511	0.760	0.671	0.485	0.542	0.475	0.574
Omni-Fuse	0.614	0.739	**0.795**	0.511	0.614	0.543	0.636
Mask-Difuser	0.628	0.756	**0.802**	0.536	0.641	0.579	0.657
CLDyN	0.646	0.773	0.764	0.552	0.658	0.621	0.669
ISFusion	0.665	**0.784**	0.736	**0.574**	**0.672**	0.684	**0.686**
Ours (CNN)	0.677	**0.813**	0.709	0.427	0.655	**0.753**	0.672
Ours (Mamba)	**0.683**	**0.791**	**0.823**	**0.591**	**0.689**	**0.712**	**0.715**
Ours (Transformer)	**0.701**	0.781	0.774	**0.633**	**0.669**	**0.822**	**0.730**

Table 5: Quantitative results of various methods for medical segmentation task on the BraTS2018 [9].

Case	T1CE	FLAIR	EMMA	Text-DiFuse	Tc-MoA
IoU [55]	0.695	0.657	0.743	0.685	0.689
Dice [22]	0.807	0.770	0.841	0.751	0.829
Case	ReFusion	C2RF	GIF-Net	CCF	Mask-Difuser
IoU	0.639	0.724	0.639	0.661	0.723
Dice	0.757	0.833	0.758	0.552	0.829
Case	CLDyN	ISFusion	Ours (CNN)	Ours (Mamba)	Ours (Transformer)
IoU	0.745	0.694	**0.782**	**0.791**	**0.799**
Dice	0.843	0.806	**0.846**	**0.854**	**0.850**

5 Conclusion

In this paper, we present a relation-constrained supervision paradigm that moves MMIF supervision beyond the spatial domain. This paradigm learns a relation space from frozen pretrained representations and aligns supervision with the goal of fusion. Our adapter aligns heterogeneous DINO and CLIP features and infers relation parameters for shared integration, complementary preservation, and conflict coordination. With adapter-tailored contrastive ranking and alternating training, the proposed space is effectively optimized. Extensive experiments validate its effectiveness.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China [No.62401097, 62601484]; Fundamental Research Funds for Central Universities, Dalian Minzu University [No.0854-53]; Liaoning Province Applied Basic Research Program [2026JH2/101300163]; Liaoning Province Science and Technology Joint Plan (2024JH2/102600113). The authors declare no competing interests.

References

[1] H. Bai, J. Zhang, Z. Zhao, Y. Wu, L. Deng, Y. Cui, T. Feng, and S. Xu. Task-driven image fusion with learnable fusion loss. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7457–7468, 2025.

[2] H. Bai, Z. Zhao, J. Zhang, Y. Wu, L. Deng, Y. Cui, B. Jiang, and S. Xu. Refusion: Learning image fusion from reconstruction with learnable loss via meta-learning. *International Journal of Computer Vision*, 133(5):2547–2567, 2025.

[3] V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca. Universeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21438–21451, 2023.

[4] B. Cao, X. Xu, P. Zhu, Q. Wang, and Q. Hu. Conditional controllable image fusion. *Advances in Neural Information Processing Systems*, 37:120311–120335, 2024.

[5] Z. Cao, Y. Zhong, Z. Wang, and L.-J. Deng. Mmaif: Multi-task and multi-degradation all-in-one for image fusion with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11744–11754, 2025.

[6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28102–28112, 2025.

[7] J. Cheng, D. Liang, and S. Tan. Transfer clip for generalizable image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25974–25984, 2024.

[8] G. Cui, H. Feng, Z. Xu, Q. Li, and Y. Chen. Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition. *Optics Communications*, 341:199–209, 2015.

[9] A. Kori, M. Soni, B. Pranjal, M. Khened, V. Alex, and G. Krishnamurthi. Ensemble of fully convolutional neural network for brain tumor segmentation from magnetic resonance images. In *International MICCAI Brainlesion Workshop*, pages 485–496. Springer, 2018.

[10] X. Li, S. Du, Y. Zou, H. Xu, Z. Jiang, and J. 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*, pages 33869–33880, 2026.

[11] X. Li, J. Liu, Z. Chen, Y. Zou, L. Ma, X. Fan, and R. Liu. Contourlet residual for prompt learning enhanced infrared image super-resolution. In *European Conference on Computer Vision*, pages 270–288. Springer, 2024.

[12] X. Li, Z. Wang, Y. Zou, Z. Chen, J. Ma, Z. Jiang, L. Ma, and J. Liu. Difiisr: A diffusion model with gradient guidance for infrared image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7534–7544, 2025.

[13] X. Li, H. Xu, S. Li, X. Chen, Z. Jiang, and J. Liu. Drfusion: Drift-resilient temporally consistent infrared-visible video fusion. *arXiv preprint arXiv:2605.25775*, 2026.

[14] X. Li, H. Xu, X. Zhu, J. Ma, Y. Zou, Z. Jiang, and J. Liu. Uncertainty-aware spatial-frequency registration and fusion for infrared and visible images. *arXiv preprint arXiv:2605.13049*, 2026.

[15] B. Liu, X. Liu, X. Jin, P. Stone, and Q. Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems*, 2021.

[16] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. 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*, pages 5802–5811, 2022.

[17] J. Liu, X. Li, Z. Wang, Z. Jiang, W. Zhong, W. Fan, and B. Xu. Promptfusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*, 2024.

[18] J. Liu, G. Wu, Z. Liu, D. Wang, Z. Jiang, L. Ma, W. Zhong, X. Fan, and R. Liu. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2349–2369, 2024.

[19] J. Liu, B. Zhang, Q. Mei, X. Li, Y. Zou, Z. Jiang, L. Ma, R. Liu, and X. Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2226–2235, 2025.

[20] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

[21] 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 Journal of Automatica Sinica*, 9(7):1200–1217, 2022.

[22] F. Milletari, N. Navab, and S.-A. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[23] A. Navon, A. Shamsian, I. Achituve, H. Maron, K. Kawaguchi, G. Chechik, and E. Fetaya. Multi-task learning as a bargaining game. In *International Conference on Machine Learning*, 2022.

[24] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[25] G. Qu, D. Zhang, and P. Yan. Information measure for performance of image fusion. *Electronics letters*, 38(7):313–315, 2002.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[27] Y.-J. Rao. In-fibre bragg grating sensors. *Measurement science and technology*, 8(4):355–375, 1997.

[28] J. W. Roberts, J. A. Van Aardt, and F. B. Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008.

[29] D. Senushkin, N. Patakin, A. Kuznetsov, and A. Konushin. Independent component alignment for multi-task learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[30] L. Tang, C. Li, and J. Ma. Mask-difuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[31] L. Tang, Q. Yan, X. Xiang, L. Fang, and J. Ma. C2rf: Bridging multi-modal image registration and fusion via commonality mining and contrastive learning. *International journal of computer vision*, 133(8):5262–5280, 2025.

[32] Y. Tian, Q. Ye, and D. Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025.

[33] P.-w. Wang and B. Liu. A novel image fusion metric based on multi-scale analysis. In *2008 9th international conference on signal processing*, pages 965–968. IEEE, 2008.

[34] X. Wang, Z. Guan, W. Qian, C. Wang, and R. Ma. Multi-modal image fusion via intervention-stable feature learning. *arXiv preprint arXiv:2603.23272*, 2026.

[35] Z. Wang, J. Zhang, T. Guan, Y. Zhou, X. Li, M. Dong, and J. Liu. Efficient rectified flow for image fusion. *Advances in Neural Information Processing Systems*, 2025.

[36] Z. Wang, L. Zhao, J. Zhang, R. Song, H. Song, J. Meng, and S. 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, 2025.

[37] G. Wu, H. Liu, H. Fu, Y. Peng, J. Liu, X. Fan, and R. Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17882–17891, 2025.

[38] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo. Fusiondn: A unified densely connected network for image fusion. In *proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[39] X. Xu, S. Kong, T. Hu, Z. Liu, and H. Bao. Boosting image restoration via priors from pre-trained models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2900–2909, 2024.

[40] C. S. Xydeas and V. Petrovic. Objective image fusion performance measure. *Electronics letters*, 36(4):308–309, 2000.

[41] 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. *arXiv preprint arXiv:2604.08924*, 2026.

[42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27026–27035, 2024.

[43] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 2020.

[44] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[45] H. Zhang, L. Cao, and J. Ma. Text-difuse: An interactive multi-modal image fusion framework based on text-modulated diffusion model. *Advances in Neural Information Processing Systems*, 37:39552–39572, 2024.

[46] H. Zhang, L. Cao, X. Zuo, Z. Shao, and J. Ma. Omnifuse: Composite degradation-robust image fusion with language-driven semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[47] H. Zhang, X. Zuo, J. Jiang, C. Guo, and J. Ma. Mrfs: Mutually reinforcing image fusion and segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26974–26983, 2024.

[48] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang. Ifcnn: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54:99–118, 2020.

[49] J. Zhao, R. Laganiere, and Z. Liu. Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement. *Int. J. Innov. Comput. Inf. Control*, 3(6):1433–1447, 2007.

[50] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5906–5916, 2023.

[51] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, K. Zhang, S. Xu, D. Chen, R. Timofte, and L. Van Gool. Equivariant multi-modality image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 25912–25921, 2024.

[52] Z. Zhao, L. Deng, H. Bai, Y. Cui, Z. Zhang, Y. Zhang, H. Qin, D. Chen, J. Zhang, P. Wang, et al. Image fusion via vision-language model. *arXiv preprint arXiv:2402.02235*, 2024.

[53] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, and J. Zhang. Didfuse: Deep image decomposition for infrared and visible image fusion. *arXiv preprint arXiv:2003.09210*, 2020.

[54] P. Zhu, Y. Sun, B. Cao, and Q. Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7099–7108, 2024.

[55] P. K. H. Zou, P. S. K. Warfield, M. A. Bharatha, M. C. M. Tempany, P. M. R. Kaus, P. S. J. Haker, P. W. M. Wells III, M. F. A. Jolesz, and M. R. Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index1: scientific reports. *Academic radiology*, 11(2):178–189, 2004.

A Pseudocode of the Training Procedure

We provide the pseudocode of the proposed training procedure in Algorithm 1 and Algorithm 2. Algorithm 1 summarizes the full optimization pipeline. It first warms up the fusion network with a conventional fusion loss to obtain initial fused images. This warm-up is only used to provide preliminary fused samples for adapter learning. After that, the fusion network is fixed and the adapter is initialized with the self-supervised ranking objective. The later training follows an alternating scheme. In each round, LFA is first updated for several epochs using the current fused images, and the fusion network is then updated for several epochs using only the relation-constrained supervision. The conventional fusion loss is not used in this alternating stage.

Algorithm 1 Training procedure of the proposed framework

Input: Paired source dataset $\mathcal{D} = \{(A, B)\}$.
Modules: Frozen feature extractor Ψ, fusion network $\mathcal{F}_\Theta$, learnable feature adapter $\mathcal{A}_\Phi$.
Initialize parameters Θ and Φ.
// Warm-up fusion network to obtain initial fused images
for epoch $= 1$ to E_{warm} **do**
  for each mini-batch $(A, B) \subset \mathcal{D}$ **do**
    $F = \mathcal{F}_\Theta(A, B)$
    Update Θ by back-propagating $\mathcal{L}_{\mathrm{warm}}(F; A, B)$
  end for
end for
// Initialize adapter with fixed fusion network
Freeze Θ.
for epoch $= 1$ to E_{ada} **do**
  for each mini-batch $(A, B) \subset \mathcal{D}$ **do**
    Sample unpaired sources A^- and B^-.
    $F = \mathcal{F}_\Theta(A, B)$
    $\mathcal{L}_{\mathrm{adapter}} = \textsc{AdapterObjective}(A, B, A^-, B^-, F; \Phi)$
    Update Φ by back-propagating $\mathcal{L}_{\mathrm{adapter}}$
  end for
end for
// Alternating optimization
for round $= 1$ to K **do**
// Adapter update: train LFA for several epochs with current fused images
  Freeze Θ and unfreeze Φ.
  for epoch $= 1$ to $E_{\mathrm{ada}}^{\mathrm{alt}}$ **do**
    for each mini-batch $(A, B) \subset \mathcal{D}$ **do**
      Sample unpaired sources A^- and B^-.
      $F = \mathcal{F}_\Theta(A, B)$
      $\mathcal{L}_{\mathrm{adapter}} = \textsc{AdapterObjective}(A, B, A^-, B^-, F; \Phi)$
      Update Φ by back-propagating $\mathcal{L}_{\mathrm{adapter}}$
    end for
  end for
// Fusion update: train FusionNet for several epochs only with RCS
  Freeze Φ and unfreeze Θ.
  for epoch $= 1$ to $E_{\mathrm{fus}}^{\mathrm{alt}}$ **do**
    for each mini-batch $(A, B) \subset \mathcal{D}$ **do**
      $F = \mathcal{F}_\Theta(A, B)$
      $\mathcal{L}_{\mathrm{rel}} = \mathrm{RCS}(A, B, F; \Phi)$
      Update Θ by back-propagating $\mathcal{L}_{\mathrm{rel}}$
    end for
  end for
end for
Output: Trained fusion network $\mathcal{F}_\Theta$ and adapter $\mathcal{A}_\Phi$.

Algorithm 2 details the adapter objective used in both adapter initialization and adapter-update stages. Given a matched source pair, two unpaired sources, and the current fused image, the adapter first

computes source-pair scores for the matched and mismatched pairs. These scores are used to form the source–source ranking loss $\mathcal{L}_{\text{ssr}}$. The fused image is then encoded by feeding the self-pair (F, F) into the same pair-conditioned adapter, producing the fused representation z_F. Based on z_F, the adapter computes one matched source–fused score $s_{F,(A,B)}$ and two mismatched source–fused scores $s_{F,(A,B^-)}$ and $s_{F,(A^-,B)}$. These scores are used to form the source–fused ranking loss $\mathcal{L}_{\text{sfr}}$, which encourages the fused image to be more compatible with its matched source pair than with mismatched source pairs. The final adapter objective is the sum of these two ranking losses. This design trains the adapter to distinguish matched source relations and to remain reliable when fused images enter the supervision space.

Algorithm 2 ADAPTEROBJECTIVE$(A, B, A^-, B^-, F; \Phi)$

Input: Matched source pair (A, B), unpaired sources A^- and B^-, fused image F.
Module: Learnable feature adapter $\mathcal{A}_\Phi$.
Margins: δ_{src}, δ_{fuse}.

Extract $\Psi(X) = \{f_D^4(X), f_C^6(X), f_D^8(X)\}$ for $X \in \{A, B, A^-, B^-, F\}$.
// Source–source ranking
$\mathcal{A}_\Phi(\Psi(A), \Psi(B)) \to (z_A^{A,B}, z_B^{A,B}, m_{A,B}, d_{A,B}, r_{A,B}) \to s_{A,B}$.
$\mathcal{A}_\Phi(\Psi(A), \Psi(B^-)) \to (z_A^{A,B^-}, z_{B^-}^{A,B^-}, m_{A,B^-}, d_{A,B^-}, r_{A,B^-}) \to s_{A,B^-}$.
$\mathcal{A}_\Phi(\Psi(A^-), \Psi(B)) \to (z_{A^-}^{A^-,B}, z_B^{A^-,B}, m_{A^-,B}, d_{A^-,B}, r_{A^-,B}) \to s_{A^-,B}$.
Compute source–source ranking loss:

$$\mathcal{L}_{\text{ssr}} = \frac{1}{2}\Big(\max(0, \delta_{\text{src}} - s_{A,B} + s_{A,B^-}) + \max(0, \delta_{\text{src}} - s_{A,B} + s_{A^-,B}) \Big). \tag{17}$$

// Matched source pair should score higher than mismatched source pairs
// Source–fused ranking
Encode the fused self-pair: $\mathcal{A}_\Phi(\Psi(F), \Psi(F)) \to z_F$.
$(z_A^{A,B}, z_B^{A,B}, m_{A,B}, d_{A,B}, r_{A,B}, z_F) \to s_{F,(A,B)}$.
$(z_A^{A,B^-}, z_{B^-}^{A,B^-}, m_{A,B^-}, d_{A,B^-}, r_{A,B^-}, z_F) \to s_{F,(A,B^-)}$.
$(z_{A^-}^{A^-,B}, z_B^{A^-,B}, m_{A^-,B}, d_{A^-,B}, r_{A^-,B}, z_F) \to s_{F,(A^-,B)}$.
Compute source–fused ranking loss:

$$\mathcal{L}_{\text{sfr}} = \frac{1}{2}\Big(\max(0, \delta_{\text{fuse}} - s_{F,(A,B)} + s_{F,(A,B^-)}) + \max(0, \delta_{\text{fuse}} - s_{F,(A,B)} + s_{F,(A^-,B)}) \Big). \tag{18}$$

// Fused image should score higher with its matched source pair than with mismatched source pairs
Compute adapter objective:

$$\mathcal{L}_{\text{adapter}} = \mathcal{L}_{\text{ssr}} + \mathcal{L}_{\text{sfr}}. \tag{19}$$

Output: $\mathcal{L}_{\text{adapter}}$.

B Problem Statement and Modeling

Given two source images A and B, the goal of MMIF is to learn a fusion network $\mathcal{F}$ that generates a fused image: $F = \mathcal{F}(A, B; \Theta)$, where Θ denotes the parameters of the fusion network. Under this notation, most existing multimodal image fusion methods follow a spatial source-image supervision paradigm, which directly uses the source images as GT to compute loss, defined as:

$$\arg\min_{\Theta} \mathcal{L}_{\text{spa}}(F; A, B). \tag{20}$$

Here, $\mathcal{L}_{\text{spa}}$ denotes the losses composed of spatial-domain criteria such as intensity, gradient, and structural similarity. However, A and B are carriers of information, rather than natural ground truth for the fused image. They typically contain shared information, modality-specific information, and potential conflicts simultaneously. Therefore, directly using spatial-domain source images as GT often drives fusion learning toward pixel-level compromise or modality bias, rather than faithfully modeling the integration objective of MMIF. To address this limitation, we argue that an ideal supervisory space for MMIF should align with the fundamental objective of fusion. Since MMIF aims to integrate shared information, preserve complementary information, and handle potential

cross-modal conflicts, we construct the proposed supervisory space from these three aspects. In other words, the proposed fusion supervision is defined not by direct spatial-domain approximation to source images, but by whether the fused result satisfies the fundamental requirements of MMIF in a knowledge-driven relational space. Under this formulation, fusion supervision is defined as:

$$\arg\min_{\Theta} \underbrace{\mathcal{L}_{\mathrm{rel}}}_{\mathcal{L}_{\mathrm{shr}},\mathcal{L}_{\mathrm{cmp}},\mathcal{L}_{\mathrm{crd}}} \Big(\underbrace{\mathcal{A}\big(\Psi(A),\Psi(B),\Psi(F);\Phi\big)}_{\hat{A},\hat{B},\hat{F},m,d,r} \Big). \tag{21}$$

Here, Φ denotes the parameters of the learnable feature adapter $\mathcal{A}$, which outputs aligned features z_A, z_B, and z_F in the unified supervision space, along with three relation parameters m, d, and r. Built on these outputs, $\mathcal{L}_{\mathrm{rel}}$ denotes the proposed relation-constrained supervision with three components: the shared loss $\mathcal{L}_{\mathrm{shr}}$, the complementary loss $\mathcal{L}_{\mathrm{cmp}}$, and the coordination loss $\mathcal{L}_{\mathrm{crd}}$.

C Why Fusion Supervision Requires Sharedness, Complementarity, and Coordination?

A proper supervision space for MMIF should reflect the objective of fusion rather than the appearance of any single source image. Given a source pair (A, B), the fused image is expected to integrate information that is jointly supported by both sources, retain useful information that appears mainly in one source, and avoid unstable updates caused by cross-modal conflicts. These requirements naturally lead to three relations: sharedness, complementarity, and coordination.

Sharedness. Sharedness describes the degree to which two sources provide consistent information at a local position. When the two sources contain the same structure or semantic content, the fused representation should not be biased toward either source. Instead, it should preserve their common information. Without modeling sharedness, a supervision loss may treat the two source images as two independent targets and force the fusion model to make a pixel-level compromise. This may weaken common structures or introduce modality bias. Therefore, sharedness is needed to identify where the fused representation should follow the common representation.

Complementarity. Complementarity describes useful non-shared information provided by different sources. In MMIF, some important cues are visible only in one modality, such as thermal targets in infrared images or fine textures in visible images. A supervision space that only emphasizes shared information would push the fused representation toward common cues and may suppress modality-specific cues. Therefore, complementarity is needed to guide the fused representation to preserve non-shared but useful information. In our formulation, this is achieved by supervising the fused offset rather than only its distance to the shared representations.

Coordination. Coordination is required because not all source differences are useful complementarity. Some regions contain strong cross-modal disagreement, noise, or inconsistent responses. If the supervision always forces the fused representation to follow one source or preserve both sources without restriction, the fusion result may be driven excessively toward one modality. This can introduce artifacts or unstable representations. Coordination defines an adaptive tolerance range for the fused offset. It allows useful complementary deviations while suppressing excessive deviations caused by conflict.

Together, the three relations define different but connected aspects of fusion supervision. Sharedness characterizes the degree of common information between the sources and modulates the strength of complementary supervision. Complementarity determines how non-shared source cues should be preserved. Coordination determines how far the fused representation can deviate under conflict. Thus, the proposed supervision space is not defined by direct spatial approximation to source images, but by whether the fused representation satisfies these three requirements induced by source relations.

D Additional Analysis of the Relation-Constrained Supervision

This appendix provides a detailed explanation of how the relation parameters affect the supervision structure, the relation-constrained losses, the relatedness scores, and the training objective. The goal is to clarify how the proposed supervision space responds to different source relations.

D.1 Effects of Relation Parameters on Supervision

Sharedness and Dominance (m_i and d_i). The sharedness parameter m_i characterizes the degree of shared information between the two modalities at token i and modulates the complementary loss through the factor $1 - m_i$. The dominance parameter d_i controls the relative contribution of the two sources in non-shared regions through the dominance-aware complementary offset:

$$v_i = d_i u_{A,i} + (1 - d_i)\, u_{B,i}. \tag{22}$$

Here, v_i denotes the dominance-aware complementary offset. When d_i is close to 1 or 0, the target offset favors source A or B, respectively, while d_i close to 0.5 gives both sources more balanced contributions. The complementary loss is defined as:

$$\mathcal{L}_{\text{cmp}} = \frac{1}{N} \sum_{i=1}^{N} (1 - m_i)\, \mathcal{D}_{\cos}(u_{F,i}, v_i). \tag{23}$$

Thus, higher sharedness weakens the complementary constraint, while lower sharedness strengthens it to preserve modality-specific cues. Meanwhile, reducing $\mathcal{L}_{\text{cmp}}$ encourages the fused offset $u_{F,i}$ to follow the dominance-aware direction in non-shared regions. The term $|2d_i - 1|$ is also used in the relatedness score. It becomes larger when the dominance relation is clearer, allowing such tokens to provide stronger relation evidence during adapter ranking.

Coordination radius r_i. The coordination radius r_i controls the allowed range of the fused offset. Its effect appears in the coordination loss:

$$\mathcal{L}_{\text{crd}} = \frac{1}{N} \sum_{i=1}^{N} \max(0, \|u_{F,i}\|_2 - r_i)\,. \tag{24}$$

When r_i is small, the fused offset is constrained to stay close to the source-induced reference. When r_i is large, the loss becomes more tolerant and allows a wider offset range. This is useful in regions with strong cross-modal conflict, where strict alignment may suppress useful information. However, a large r_i also indicates that the local relation is less stable. Therefore, the relatedness score reduces the contribution of such tokens through the negative term $-\lambda_r r_i$ in the relation-aware weight. In this way, r_i weakens overly strict supervision in difficult regions while preventing unstable tokens from dominating adapter ranking.

E Why Self-Supervised Contrastive Learning Works for LFA?

LFA must learn pair-conditioned representations and relation parameters before it can supervise the fusion network. Since MMIF has no ground-truth fused image, these outputs cannot be trained with direct target labels. We instead use the pairing structure of source images as a self-supervised signal. For a matched source pair (A, B), the two images are captured from the same scene and share objects, structures, and semantic layout. For mismatched pairs such as (A, B^-) and (A^-, B), this scene-level correspondence is broken. This provides a natural contrastive signal for adapter learning.

E.1 Source-Side Contrastive Signal

The source-side signal trains LFA to distinguish matched source pairs from mismatched ones. Instead of requiring an absolute ground-truth relation score, it only requires the matched pair to have a higher relatedness score:

$$s_{A,B} > s_{A,B^-}, \qquad s_{A,B} > s_{A^-,B}. \tag{25}$$

This relative constraint is suitable for LFA because the adapter is expected to learn a relation space, rather than predict a fixed label for each pair. The corresponding source–source ranking loss is defined as:

$$\mathcal{L}_{\text{ssr}} = \frac{1}{2}\Big(\max(0, \delta_{\text{src}} - s_{A,B} + s_{A,B^-}) + \max(0, \delta_{\text{src}} - s_{A,B} + s_{A^-,B}) \Big). \tag{26}$$

The margin δ_{src} prevents weak separation. Thus, LFA cannot satisfy the objective by assigning similar scores to all pairs. It must learn representations and relation parameters that separate matched and mismatched source relations.

E.2 Coupled Training of Representations and Relation Parameters

The relatedness score is designed to depend on both types of LFA outputs. For a source pair (A, B), the score is defined as:

$$s_{A,B} = \frac{\sum_{i=1}^{N} w_i \cos(z_{A,i}, z_{B,i})}{\sum_{i=1}^{N} w_i + \epsilon}, \quad w_i = \sigma(\lambda_m m_i + \lambda_d |2d_i - 1| - \lambda_r r_i)\,. \tag{27}$$

The cosine term trains the pair-conditioned representations z_A and z_B. It requires matched source pairs to be locally consistent and mismatched pairs to be less consistent. The weight w_i trains the relation parameters. A larger m_i gives higher weight to shared regions, and a larger $|2d_i - 1|$ gives higher weight to regions with clearer dominance. A larger r_i reduces the weight because a wider coordination range indicates a less stable local relation.

This design prevents one part of LFA from being optimized alone. If z_A and z_B are not discriminative, matched and mismatched pairs cannot be separated. If m_i, d_i, and r_i do not identify reliable relation evidence, unstable tokens may dominate the score. Therefore, the ranking objective jointly calibrates the pair-conditioned representations and the relation parameters through the same score function.

E.3 Source-Fused Contrastive Signal

Source-side ranking alone is insufficient because LFA is later used to supervise fused images. If the adapter only sees source-source pairs during training, it may not remain reliable when a fused image enters the supervision space. To reduce this mismatch, we include the current fused image F in adapter training and impose the source–fused ranking constraint:

$$s_{F,(A,B)} > s_{F,(A^-,B)}, \qquad s_{F,(A,B)} > s_{F,(A,B^-)}. \tag{28}$$

Here, $s_{F,(A,B)}$ measures the compatibility between the fused representation and the relation-aware reference induced by its matched source pair, while $s_{F,(A^-,B)}$ and $s_{F,(A,B^-)}$ measure its compatibility with references induced by mismatched source pairs. A valid fused representation should integrate the shared and complementary information from its corresponding sources. Therefore, it should better match its reference from the matched source pair than those from mismatched pairs.

The source–fused score follows the same structure as the relation-constrained supervision:

$$s_{F,(A,B)} = \frac{\sum_{i=1}^{N} w_i \Big(m_i c_{shr} + (1 - m_i) \cos\big(u_{F,i},\, d_i u_{A,i} + (1 - d_i) u_{B,i}\big) \Big)}{\sum_{i=1}^{N} w_i + \epsilon}. \tag{29}$$

where $c_{shr} = \cos(c_{F,i}, c_{A,i}) + \cos(c_{F,i}, c_{B,i})$. The first term checks whether $z_{F,i}$ is close to the shared cues in shared regions. The second term checks whether $u_{F,i}$ follows the dominance-aware complementary offset in non-shared regions. Thus, source–fused ranking does not simply add fused images as extra samples. It trains LFA on the same type of relation that is later used for fusion supervision.

The corresponding source–fused ranking loss is defined as:

$$\mathcal{L}_{\text{sfr}} = \frac{1}{2}\Big(\max(0, \delta_{\text{fuse}} - s_{F,(A,B)} + s_{F,(A,B^-)}) + \max(0, \delta_{\text{fuse}} - s_{F,(A,B)} + s_{F,(A^-,B)}) \Big). \tag{30}$$

Here, the margin δ_{fuse} plays the same role as δ_{src}, but it is applied to source–fused ranking. The full adapter objective is:

$$\mathcal{L}_{\text{adapter}} = \mathcal{L}_{\text{ssr}} + \mathcal{L}_{\text{sfr}}. \tag{31}$$

E.4 Sufficient Pretraining and Alternating Optimization

The adapter receives supervision from both pair discrimination and fused-image consistency. For each matched source pair, the source–source ranking loss compares it with two mismatched pairs. This trains LFA to identify reliable source relations. The source–fused ranking loss then introduces the current fused image into adapter learning. Since both scores depend on z and the relation parameters m, d, r, these objectives jointly update the representation space and the relation parameters.

The warm-up fusion network provides the initial fused images required by source–fused ranking. These fused images do not need to be optimal. They only provide a starting distribution for adapting LFA from source-source relations to source-fused relations. After warm-up, LFA is trained with the

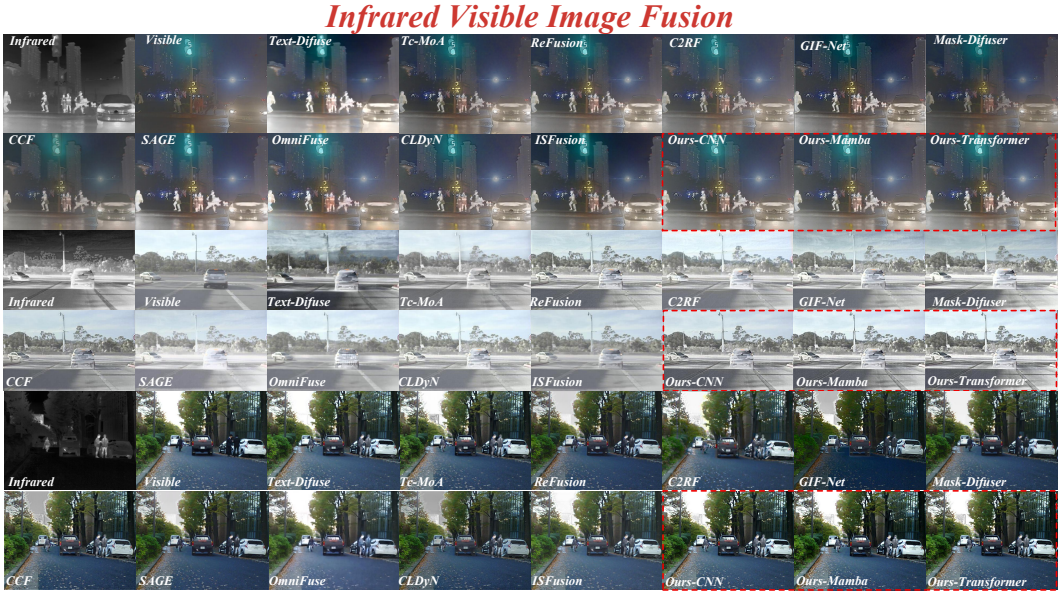


Figure 7: Full qualitative comparison for IVF task on M^{3}FD, RoadScene and MSRS.

fusion network fixed. The current fused image F is generated by the fixed fusion network and is used in $\mathcal{L}_{\text{sfr}}$.

The following training proceeds alternately. During the adapter-update stage, the fusion network is frozen, and LFA is updated with $\mathcal{L}_{\text{adapter}}$ using the current fused images. During the fusion-update stage, LFA is frozen, and the fusion network is optimized only with $\mathcal{L}_{\text{rel}}$. No conventional fusion loss is used in this stage. This creates a closed training loop: improved fused images provide more informative source–fused samples for LFA, and an improved LFA provides a more reliable supervision space for the fusion network.

F Full Qualitative Comparisons

We provide full qualitative comparisons for both MIF and IVF tasks. For MIF, the results are reported on CT-MRI [45], MRI-PET [45], and MRI-SPECT [45] fusion. For IVF, the results are reported on M^{3}FD [16], RoadScene [38], and MSRS [21]. Since our supervision paradigm is independent of the fusion architecture, we include CNN-, Mamba-, and Transformer-based variants in all comparisons.

As shown in Fig. 8, our method produces visually reliable medical fusion results across all three MIF settings. In CT-MRI fusion, our variants preserve clear anatomical structures while maintaining soft-tissue contrast from MRI. In MRI-PET and MRI-SPECT fusion, our results retain functional color information and avoid excessive color bleeding or structural distortion. These results show that the proposed supervision can guide different backbones to preserve complementary medical information in a stable manner.

Fig. 7 further shows the full qualitative comparison on IVF datasets. Existing methods may suffer from weak thermal targets, over-smoothed textures, or insufficient visible-detail preservation in challenging night and road scenes. In contrast, our CNN, Mamba, and Transformer variants consistently preserve salient infrared targets while keeping visible structures, such as vehicles, pedestrians, roads, and background textures. This indicates that the proposed relation-constrained supervision is not tied to a specific backbone and can provide effective guidance across different fusion architectures.

G Details of Source-Preservation Update Conflict Analysis

We provide the detailed procedure for the source-preservation update conflict analysis used in Table 2. Following task-gradient conflict analysis in multi-task optimization [43, 15, 23, 29], we use the update directions induced by two source-preservation objectives to examine whether a fused result

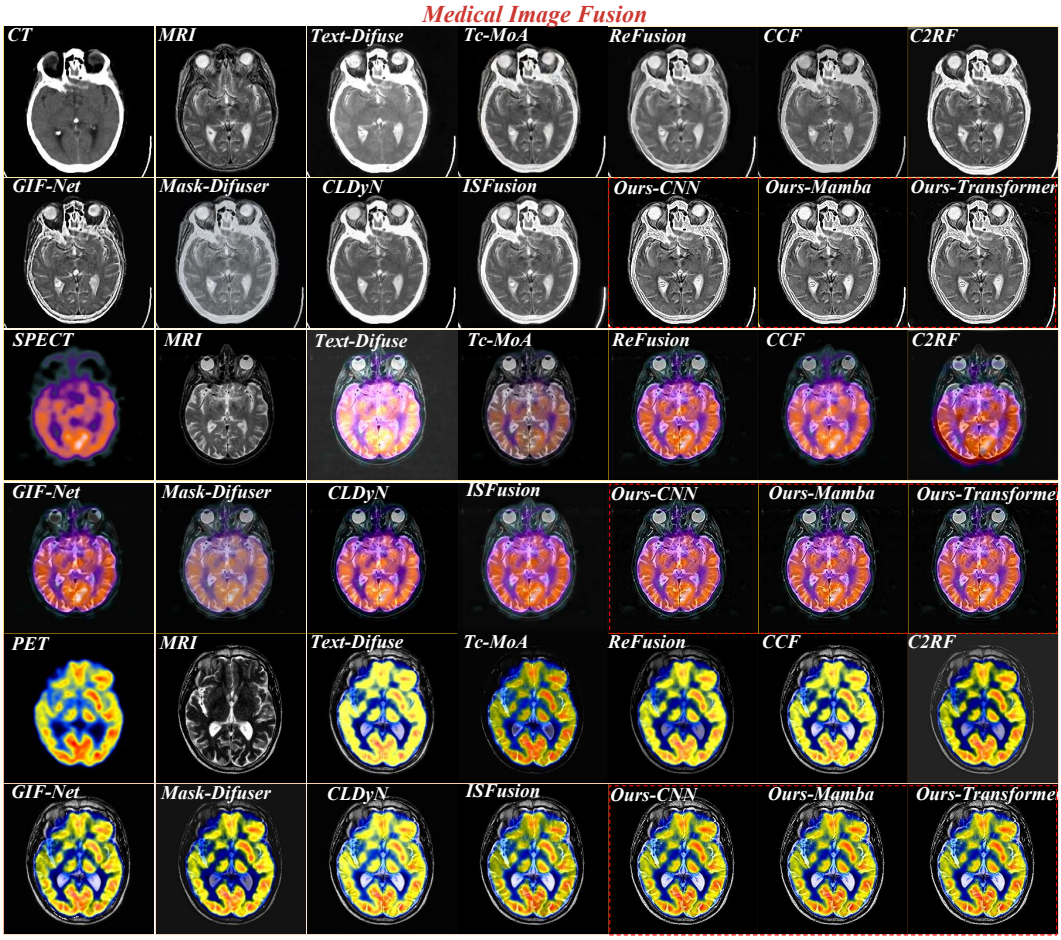


Figure 8: Full qualitative comparison for MIF task on CT-MRI, MRI-PET and MRI-SPECT.

can accommodate infrared and visible information without strong local conflict. This analysis is only used for evaluation and is not used to train any model.

For a fused result F^m produced by method m, we define the source-preservation objective for source $s \in \{ir, vis\}$ as:

$$\mathcal{L}_s^{sp} = \|F^m - I_s\|_1 + \lambda_g \left(\|D_x F^m - D_x I_s\|_1 + \|D_y F^m - D_y I_s\|_1\right),$$

where I_s denotes the source image, and D_x, D_y are Sobel operators. Here, $\|\cdot\|_1$ denotes the mean absolute error over pixels. The induced update direction is:

$$u_s^m = -\frac{\partial \mathcal{L}_s^{sp}}{\partial F^m}.$$

The negative sign converts the loss gradient into the local descent direction. Since both source-preservation gradients are negated, the cosine-based conflict criterion is unchanged compared with using the original loss gradients.

For each patch p, we restrict the two update maps to this patch and flatten them into vectors, denoted as $u_{ir,p}^m$ and $u_{vis,p}^m$. Their directional agreement is:

$$\rho_p^m = \frac{\langle u_{ir,p}^m, u_{vis,p}^m \rangle}{\|u_{ir,p}^m\|_2 \|u_{vis,p}^m\|_2 + \epsilon}.$$

A negative ρ_p^m means that infrared and visible preservation require opposite local updates for the same fused result.

High-disagreement patch set Ω. We evaluate the conflict on high-disagreement regions, where the two source images are more likely to contain competing information. For each patch p, we compute a source discrepancy score:

$$d_p = \frac{1}{|p|} \sum_{q \in p} \left(|I_{ir}^q - I_{vis}^q| + |D_x I_{ir}^q - D_x I_{vis}^q| + |D_y I_{ir}^q - D_y I_{vis}^q| \right).$$

The high-disagreement patch set Ω is defined as the top 30% patches ranked by d_p. Since Ω is computed only from the two source images, it is fixed for each image pair and shared by all compared methods.

Conflict ratio. The conflict ratio is defined as:

$$\mathrm{CR} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \mathbb{I}(\rho_p^m < 0).$$

CR measures how frequently opposite source-preservation updates occur. A lower CR means that fewer local regions require conflicting corrections to preserve the two sources.

Conflict strength. The conflict strength is defined as:

$$\mathrm{CS} = \frac{1}{|\Omega|} \sum_{p \in \Omega} [-\rho_p^m]_+.$$

CS measures how severe the conflicts are. Unlike CR, which only counts whether a conflict exists, CS considers the magnitude of negative alignment. A value close to 1 indicates nearly opposite update directions, while a value close to 0 indicates weak or no conflict.

Update imbalance. The update imbalance is defined as:

$$\mathrm{UI} = \frac{1}{|\Omega|} \sum_{p \in \Omega} \frac{\left| \|u_{ir,p}^m\|_2 - \|u_{vis,p}^m\|_2 \right|}{\|u_{ir,p}^m\|_2 + \|u_{vis,p}^m\|_2 + \epsilon}.$$

UI is not a cosine-based conflict measure. It measures the normalized magnitude gap between the infrared and visible source-preservation updates. This metric checks whether a low conflict score is caused by suppressing one modality. A lower UI indicates that the two source-preservation updates have more balanced magnitudes.

Together, CR, CS, and UI provide complementary evidence. CR measures how often conflicts occur, CS measures how strong these conflicts are, and UI checks whether conflict reduction is achieved without suppressing one source-preservation update.

H Broader Impact of the Proposed Paradigm

Table 6 evaluates whether our paradigm can serve as a general refinement strategy for existing fusion models. For a fair plug-in evaluation, we keep the original network architecture, training dataset, and training settings of each baseline unchanged. The only modification is to replace its original supervision with our relation-constrained supervision paradigm. We apply this setting to three representative baselines, including ReFusion, EMMA, and C2RF, and report the results on the RoadScene dataset.

The upgraded versions consistently improve all seven metrics, showing that the gains are not tied to a specific backbone. EMMA* obtains large improvements on EN, Q_M, and Q_P, while ReFusion* and C2RF* also achieve clear gains on structural and perceptual quality metrics. These results indicate that relation-constrained supervision provides a transferable optimization signal. It can enhance source preservation, structural fidelity, and fusion quality across different model designs. Therefore, the proposed paradigm has broader applicability beyond our own architecture and can improve existing image fusion systems with limited modification.

I Extended Ablation Study

We provide the full ablation results in Table 7. Besides the key variants reported in the main paper, we further analyze adaptive relation parameters and adapter training choices. These additional results show consistent trends on both IVF and MIF.

Table 6: Refinement effects of our paradigm on the RoadScene dataset. Original vs. upgraded models (* denotes our paradigm inserted).

Method	$AG\uparrow$	$EN\uparrow$	$SD\uparrow$	$MI\uparrow$	$Q_{AB/F}\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$
ReFusion [2]	5.357	59.336	48.229	2.726	0.437	0.566	0.397
ReFusion*	5.692	60.108	49.356	2.916	0.523	0.612	0.418
Gain	6.3%↑	1.3%↑	2.3%↑	7.0%↑	19.7%↑	8.1%↑	5.3%↑
EMMA [51]	4.779	49.127	46.777	2.809	0.436	0.432	0.355
EMMA*	5.514	58.742	49.108	2.943	0.518	0.589	0.407
Gain	15.4%↑	19.6%↑	5.0%↑	4.8%↑	18.8%↑	36.3%↑	14.6%↑
C2RF [31]	4.989	52.270	45.576	2.720	0.519	0.565	0.388
C2RF*	5.603	58.196	48.962	2.935	0.548	0.606	0.412
Gain	12.3%↑	11.3%↑	7.4%↑	7.9%↑	5.6%↑	7.3%↑	6.2%↑

(a) Effect of adaptive relation parameters. We examine whether the relation parameters in RCS should be adaptively inferred through three variants: (1) fixing the sharedness, (2) fixing dominance, and (3) fixing the coordination radius. Results in Table 7 show that all three variants underperform the full model. Fixed sharedness weakens the distinction between shared and non-shared information. Fixed dominance prevents the model from adapting to source-specific contributions. Fixed radius reduces the flexibility of conflict coordination. These results show that sharedness, dominance, and coordination radius should be inferred from each source pair.

(b) Effect of relation-aware scoring. We evaluate the token weighting strategy by replacing the relation-aware weight w_i with uniform weights. This variant treats all tokens equally during adapter ranking and consistently degrades performance. The result indicates that not all regions contribute equally to relation learning. Tokens with reliable sharedness, clear dominance, and smaller coordination uncertainty should receive higher weights. This verifies the necessity of relation-aware scoring in the adapter objective.

(c) Effect of adapter pretraining. We study the role of adapter pretraining by directly starting alternating optimization from an untrained adapter. This setting leads to a clear performance drop on both IVF and MIF. Since the adapter defines the supervision space, an unstable adapter can provide noisy relation constraints to the fusion network. Pretraining gives LFA a discriminative initialization before it is used for fusion supervision.

(d) Effect of fixed adapter. We further test a fixed-adapter variant, where LFA is pretrained once and then kept frozen during fusion training. This variant performs worse than the full model. The result shows that a static supervision space is insufficient because fused images evolve as the fusion network improves. Updating LFA during alternating optimization allows the supervision space to adapt to current fused results.

(e) Effect of joint optimization. We also compare alternating optimization with joint optimization, where LFA and the fusion network are updated simultaneously. Joint optimization gives lower results than the full model. This suggests that changing the supervision space and the fused output at the same time can introduce unstable mutual feedback. Alternating optimization is more stable because one side is fixed while the other is updated.

J Model Complexity and Overhead Analysis

Table 8 compares the runtime, computational cost, and parameter size of representative IVF and MIF methods. Overall, our variants maintain a compact model scale and low computational overhead across both tasks. This indicates that the proposed framework does not rely on a heavy fusion backbone to achieve effective fusion performance.

Compared with many fusion methods, our models require fewer parameters and lower computation while keeping competitive inference speed. The three variants also provide different efficiency-performance trade-offs: the Mamba variant is the most lightweight, the CNN variant offers a balanced design, and the Transformer variant introduces a moderate increase in cost for stronger representation

Table 7: Extended ablation study of the proposed method on the IVF RoadScene and MIF MRI-PET datasets. This table includes additional variants for fixed relation parameters and adapter training strategies. The best results are highlighted in **bold**.

Group	ID	Description	IVF on RoadScene							MIF on MRI-PET						
			$AG\uparrow$	$EN\uparrow$	$SD\uparrow$	$MI\uparrow$	$Q_{AB/F}\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$	$AG\uparrow$	$EN\uparrow$	$SD\uparrow$	$MI\uparrow$	$Q_{AB/F}\uparrow$	$Q_M\uparrow$	$Q_P\uparrow$
(a)	(1)	w/o LFA	5.247	58.932	47.286	2.861	0.506	0.589	0.372	8.721	94.863	72.418	2.321	0.552	0.196	0.414
	(2)	w/o FAM	5.516	60.217	49.103	3.041	0.537	0.618	0.401	9.184	97.326	75.861	2.512	0.594	0.218	0.452
	(3)	w/o RRM	5.438	59.846	48.624	2.982	0.529	0.611	0.394	9.062	96.814	75.104	2.463	0.586	0.213	0.444
	(4)	w/o RCS	5.196	58.714	47.635	2.803	0.497	0.581	0.363	8.604	94.217	71.936	2.276	0.541	0.189	0.405
(b)	(1)	w/o $\mathcal{L}_{\text{shr}}$	5.642	60.384	49.427	3.086	0.545	0.626	0.410	9.346	97.942	76.412	2.574	0.607	0.224	0.461
	(2)	w/o $\mathcal{L}_{\text{cmp}}$	5.471	60.092	48.891	2.947	0.523	0.613	0.392	9.118	97.156	75.386	2.438	0.581	0.216	0.441
	(3)	w/o $\mathcal{L}_{\text{crd}}$	5.573	60.263	49.018	3.018	0.532	0.607	0.386	9.241	97.624	75.927	2.496	0.592	0.211	0.433
	(4)	fixed sharedness	5.589	60.338	49.214	3.052	0.539	0.621	0.404	9.281	97.735	76.103	2.536	0.601	0.221	0.456
	(5)	fixed dominance	5.492	60.041	48.776	2.963	0.526	0.614	0.395	9.146	97.208	75.542	2.455	0.584	0.217	0.445
	(6)	fixed radius	5.601	60.171	49.006	3.011	0.533	0.609	0.388	9.265	97.531	75.818	2.487	0.590	0.212	0.436
(c)	(1)	w/o $\mathcal{L}_{\text{ssr}}$	5.324	59.463	48.102	2.902	0.514	0.598	0.379	8.891	95.842	73.864	2.384	0.567	0.204	0.425
	(2)	w/o $\mathcal{L}_{\text{sfr}}$	5.386	59.721	48.406	2.918	0.517	0.602	0.381	8.976	96.214	74.293	2.406	0.573	0.207	0.429
	(3)	uniform w_i	5.538	60.064	48.957	3.006	0.531	0.612	0.393	9.203	97.382	75.674	2.481	0.589	0.215	0.442
	(4)	w/o adapter pretraining	5.291	59.318	47.946	2.884	0.510	0.593	0.376	8.802	95.417	73.218	2.358	0.561	0.201	0.421
	(5)	fixed adapter	5.452	59.984	48.713	2.971	0.526	0.609	0.389	9.087	96.936	75.021	2.451	0.583	0.212	0.438
	(6)	joint optimization	5.408	59.762	48.531	2.943	0.521	0.604	0.384	9.014	96.573	74.612	2.427	0.578	0.209	0.432
Full Model (Transformer)			**5.861**	**61.069**	**50.581**	**3.179**	**0.564**	**0.643**	**0.425**	**9.690**	**99.031**	**78.072**	**2.667**	**0.627**	**0.235**	**0.481**

Table 8: Comparison of runtime, computational cost (GFLOPs), and model size (Params) on IVF and MIF tasks. All values are averaged per image.

IVF				**MIF**			
Method	**Time (s)**	**GFLOPs (G)**	**Params (M)**	**Method**	**Time (s)**	**GFLOPs (G)**	**Params (M)**
EMMA	0.058	8.861	1.516	EMMA	0.009	8.861	1.516
Text-DiFuse	23.818	2742.5	119.460	Text-DiFuse	24.424	2742.5	119.460
Tc-MoA	0.543	61.000	340.580	Tc-MoA	0.161	61.000	340.580
ReFusion	0.026	98.344	1.486	ReFusion	0.012	98.344	1.486
C2RF	0.089	123.869	1.325	C2RF	0.112	123.869	1.325
GIF-Net	0.077	39.814	0.823	GIF-Net	0.084	39.814	0.823
SAGE	0.020	29.285	0.136	CCF	31.876	1114.000	552.810
Omni-Fuse	2.614	172.923	78.320	Mask-Difuser	0.503	995.706	171.262
CLDyN	0.272	174.061	0.46	CLDyN	0.311	174.06	0.46
ISFusion	0.382	273.472	1.372	ISFusion	0.401	273.472	1.372
Ours(CNN)	0.013	3.284	0.067	Ours(CNN)	0.010	3.284	0.067
Ours(Mamba)	0.008	2.137	0.038	Ours(Mamba)	0.006	2.137	0.038
Ours(Transformer)	0.028	6.912	0.142	Ours(Transformer)	0.023	6.912	0.142

ability. These results show that the proposed relation-constrained supervision can work effectively with lightweight architectures and is suitable for practical fusion scenarios where efficiency matters.

K Implementation Details

Feature extraction and adapter. We use frozen DINO and CLIP as feature providers, denoted by Ψ. For each input image X, we extract multi-level features from DINO layer 4, CLIP layer 6, and DINO layer 8, denoted by $f_D^4(X)$, $f_C^6(X)$, and $f_D^8(X)$, respectively. Each feature map is projected by a 1×1 convolution, followed by normalization and bilinear resizing. The corresponding projection operators are denoted by P_D^4, P_C^6, and P_D^8. The projected features are resized to a common token grid of 16×16 and projected to 256 channels. They are then flattened into token sequences and concatenated along the channel dimension. The feature aligner G_{align} is implemented as an MLP with hidden dimension 512 and output dimension 256, producing the aligned representation y_X.

In the relation reasoning module, the pair encoder E_{pair} takes $[y_A, y_B, |y_A - y_B|, y_A \odot y_B]$ as input and produces the pair feature $q_{A,B}$. The source-specific heads H_A and H_B produce the pair-conditioned source representations z_A and z_B. The prediction heads h_m, h_d, and h_r estimate sharedness $m_{A,B}$, dominance $d_{A,B}$, and coordination radius $r_{A,B}$, respectively. The pair encoder is implemented as an MLP with GELU activation and LayerNorm, and the prediction heads are lightweight linear layers. Sharedness and dominance are normalized by the sigmoid function, while the coordination radius is constrained by the softplus function.

Hyperparameter settings. For the relation-aware source-pair score, we set the weighting coefficients to $\lambda_m = 1.0$, $\lambda_d = 0.5$, and $\lambda_r = 0.5$. The numerical stability constant is set to $\epsilon = 1 \times 10^{-6}$. The

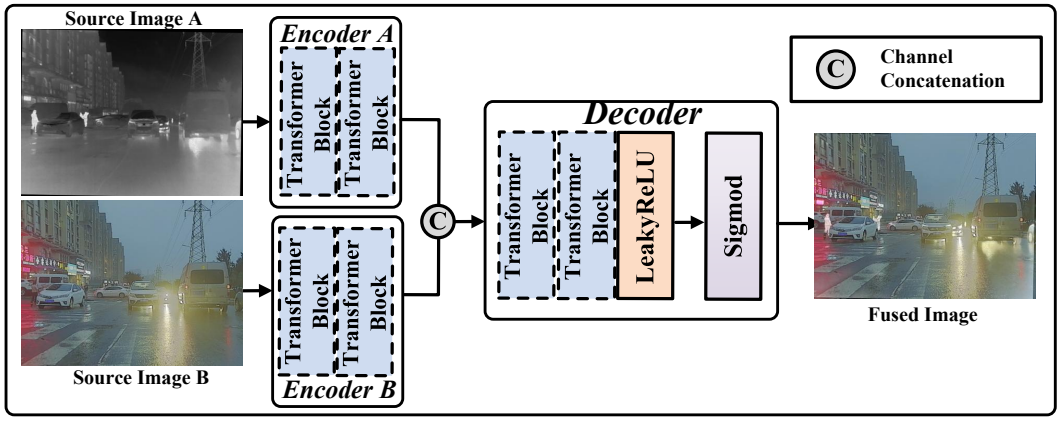


Figure 9: Schematic diagram of fusion network.

source-side ranking margin and fused-side ranking margin are set to $\delta_{\text{src}} = 0.2$ and $\delta_{\text{fuse}} = 0.2$, respectively. For the source-preservation update conflict analysis, the gradient weight is set to $\lambda_g = 1.0$. We use patch size 16×16 and define the high-disagreement region Ω as the top 30% patches ranked by source intensity and gradient discrepancy. α_1 and α_2 in $\mathcal{L}_{\text{warm}}$ are set to 1 and 2.

Fusion network architecture. As shown in Fig. 9, our FusionNet adopts a two-branch encoder-decoder architecture. The two source images are first processed by two modality-specific encoders, denoted as Encoder A and Encoder B. These two encoders have the same architecture but do not share weights, allowing each branch to learn source-specific feature extraction. Taking the Transformer-based backbone as an example, each encoder is composed of stacked Transformer blocks and extracts a feature from its corresponding source image. The two latent features are then concatenated along the channel dimension and fed into the decoder. The decoder aggregates the concatenated feature through stacked Transformer blocks and a lightweight output head with LeakyReLU and Sigmoid, producing the final fused image.

Under this unified design, we instantiate Transformer- [44], CNN- [48], and Mamba-based [20] fusion backbones to verify that the proposed supervision paradigm is not tied to a specific architecture. The Transformer variant is illustrated in Fig. 9. The CNN and Mamba variants follow the same two-branch encoder-decoder pipeline, and their only difference is that the Transformer blocks are replaced by CNN blocks [48] and VMamba blocks [20], respectively.

Architecture and Pretraining of the Shared Token Encoder. G_{shr} is implemented as a lightweight token-wise MLP with shared weights across all modalities. Given an aligned representation $z \in \mathbb{R}^{N\times C}$, each token is processed independently as:

$$G_{\text{shr}}(z) = W_2\,\phi(W_1\,\text{LN}(z))\,, \tag{32}$$

where $\text{LN}(\cdot)$ denotes LayerNorm, $\phi(\cdot)$ denotes GELU, and W_1 and W_2 are linear projections. The output dimension is kept identical to that of z, allowing the shared representation to remain in the same supervision space. The same G_{shr} is applied to the two source representations and the fused representation. The corresponding residual representations u_A and u_B are defined by direct subtraction.

Following the correlation-driven feature decomposition strategy of CDDFuse [50], we pretrain G_{shr} using matched source pairs only. We encourage the shared representations c_A and c_B to be highly correlated, while reducing the correlation between the corresponding residual representations u_A and u_B. We define the correlation between two token representations as:

$$\text{CC}(x, y) = \frac{\langle x - \mu_x,\, y - \mu_y\rangle}{\|x - \mu_x\|_2 \|y - \mu_y\|_2 + \epsilon}, \tag{33}$$

where μ_x and μ_y denote the corresponding mean representations. Following this correlation-driven decomposition principle, the pretraining objective is defined as:

$$\mathcal{L}_G = 1 - \text{CC}(c_A, c_B) + \text{CC}^2(u_A, u_B), \tag{34}$$

where the first term encourages G_{shr} to retain information consistently represented in both modalities, while the second term suppresses correlated information in the residual components. After pretraining, the parameters of G_{shr} are frozen. The same encoder is subsequently applied to z_A, z_B, and z_F during relation-constrained supervision and source–fused ranking.

L Limitations

Although our method achieves strong fusion performance with a compact model design, it still has some limitations. First, the relation-constrained supervision relies on pretrained vision models to construct the supervision space. Its effectiveness may therefore be influenced by the representation quality and domain coverage of the selected pretrained models. Second, the current design mainly focuses on pairwise source relations. Extending it to more complex fusion settings with more than two modalities may require additional relation modeling. Future work will explore more robust supervision construction and broader adaptation to multi-modal and multi-source fusion scenarios.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See our abstract.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: Our paper does not include theoretical results

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide full details for reproducing the proposed method.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our code and data will be publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide complete training and testing settings in Section 4 and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: See it in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conducted in this paper complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See it in appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: There is no such risks in our paper.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In Section 4, we introduce all datasets and metrics we used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.