DECOMPOSED AND DPP-GUIDED PARADIGM FOR INFRARED AND VISIBLE VIDEO FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Infrared and Visible Video Fusion (IVVF) integrates complementary information from heterogeneous sensors to produce informative and temporally coherent videos. However, existing methods suffer from spatio-temporal optimization imbalances, yielding motion ghosting or temporal incoherence, while inadequate cross-modal exploitation causes spatial misalignment. To address these, we propose D^2Fusion, a Decomposed and Determinantal Point Process (DPP)-guided fusion network. With better prior understanding provided by large-scale visual models, D^2Fusion enhances spatio-temporal consistency, leveraging them to modulate global motion and refine local details. Moreover, the opposing constraints between temporal consistency and spatial richness are abstracted as a quality–diversity trade-off unified under a DPP objective to balance them. Beyond intra-modal priors, we enforce a modality-invariant prior over shared geometric structures against cross-modal misalignment, preserving modality-specific cues. Extensive experiments on M3SVD establish D^2Fusion as SOTA with strong cross-dataset generalization.

1 INTRODUCTION

Multimodal video fusion (Wang et al., 2025; Han et al., 2026) aims to synthesize informative outputs by integrating complementary data from heterogeneous sensors, which is essential for robust visual perception under challenging conditions, including low-light environments (Zhang et al., 2026), adverse weather (Li et al., 2026a), and occluded scenes (Gong et al., 2025). Infrared and Visible Video Fusion (IVVF) aims to effectively integrate complementary information from infrared and visible sensor streams. The resulting fused video combines the salient target information captured by infrared sensors with the rich background details provided by visible sensors, thereby enhancing the performance of downstream tasks (Liu et al., 2024b; Wang et al., 2025).

Despite these advances, existing methods still face two fundamental limitations. One major limitation is spatio-temporal imbalance (Qiang et al., 2026; Li et al., 2026b). As shown in Fig. 1(a), temporal consistency and spatial fidelity are often competing objectives, and existing methods tend to prioritize one at the expense of the other. Overemphasizing either side leads to motion ghosting (Tang et al., 2026) or abrupt brightness changes (Cheng et al., 2025). Additionally, existing methods suffer from insufficient cross-modal information exploitation (Liu et al., 2024b; Ding et al., 2026). Content-agnostic designs prevent them from fully leveraging complementary information of different modalities, as illustrated in Fig. 1(b). Consequently, discriminative information from both modalities is not adequately preserved, and spatial misalignment arises between them.

To balance spatio-temporal optimization, we propose a decomposed-structure and unified-supervision paradigm. Architecturally, a temporal-spatial decomposed dual-branch design assigns each prior to its dedicated branch, adopting a Video Foundation Temporal Prior (VFTP) from a world model and a Video Foundation Spatial Prior (VFSP) from a ViT-based MAE. Since large-scale pre-training equips VFTP with patch-level global temporal priors and ViT-based MAE with intra-patch pixel-level priors, the decomposition leverages each prior to enhance global temporal modulation and local spatial refinement respectively. Moreover, we abstract the opposing constraints of temporal consistency and spatial richness on the same latent representation as a quality-diversity trade-off unified under a Determinantal Point Process (DPP) objective, balancing the two aspects adaptively. Beyond intra-modal priors, we enforce a modality-invariant prior over shared geometric structure

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

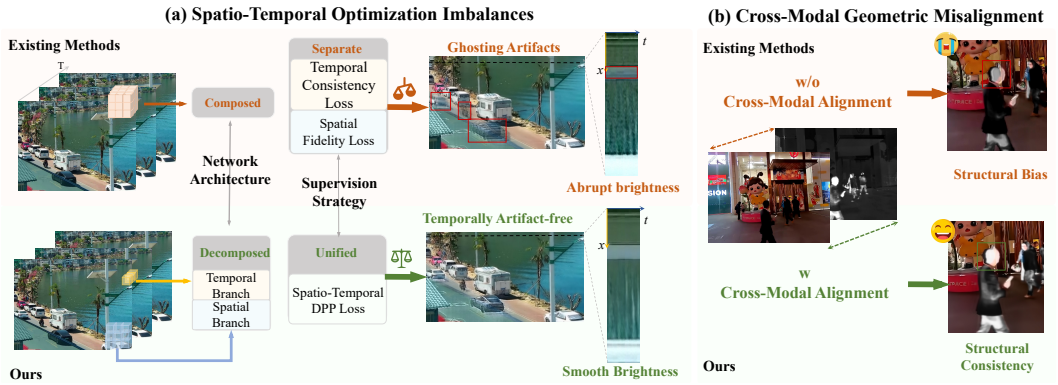

Figure 1: (a) Overemphasizing temporal consistency causes motion ghosting in dynamic regions, while overemphasizing spatial contrast causes abrupt brightness changes that degrade temporal coherence. (b) Cross-modal alignment effectively mitigates structural bias along the spatial axis.

against cross-modal misalignment, retaining modality-shared geometry while allowing modality-specific cues such as infrared thermal saliency and visible-light background context to specialize. This enhances cross-modal discrimination and the mutual information between the fused output and each source modality. Our main contributions are as follows:

- We propose D^2Fusion, a dual-stream framework that decomposes spatial and temporal feature learning via complementary priors from VFPs, unified through a DPP to address spatio-temporal optimization imbalances.
- We distill cross-modal shared geometric structure as a modality-invariant prior against cross-modal misalignment while allowing modality-specific cues to specialize.
- Our model achieves state-of-the-art performance on the M3SVD video fusion benchmark across multiple metrics, exhibits strong generalization in cross-dataset evaluation.

2 Related Work

2.1 Video Fusion

Early video fusion methods largely adapt image fusion techniques by applying per-frame strategies to video sequences, evolving from architecture-driven designs e.g. AutoEncoders (Li & Wu, 2018), CNNs (Liu et al., 2024a), Transformers (Li & Wu, 2024; Ma et al., 2022) toward semantic-driven strategies (Wu et al., 2025; Palladin et al., 2024). However, operating solely on static frames, they overlook temporal cues in videos. To exploit inter-frame complementarity, Ding et al. (2026); Peng et al. (2025) model cross-frame cues as temporal consistency constraints. Tang et al. (2026); Zhao et al. (2026) incorporate temporal information via frame-by-frame strategies, temporal losses, or recurrent and Transformer-based architectures.

2.2 Pretrained Priors for Video Tasks

Xie et al. (2024b) exploits intermediate frame as temporal evolution priors in video restoration. Drawing on this insight, recent large-scale pretrained models with strong spatio-temporal representations offer promising priors for video fusion. These two paradigms are complementary: pixel-level reconstruction (Tong et al., 2022) excels at preserving spatial structure, while latent-space feature prediction (Assran et al., 2025; Huang, 2026) is effective at modeling temporal dynamics. These priors naturally motivate structurally decomposed fusion frameworks. Beyond static spatio-temporal representations, optical flow estimators (Teed & Deng, 2020; Xu et al., 2022; Wang et al., 2024) provide cross-frame motion priors through warping, thereby facilitating temporal alignment. This work investigates how such pretrained priors can enhance video fusion.

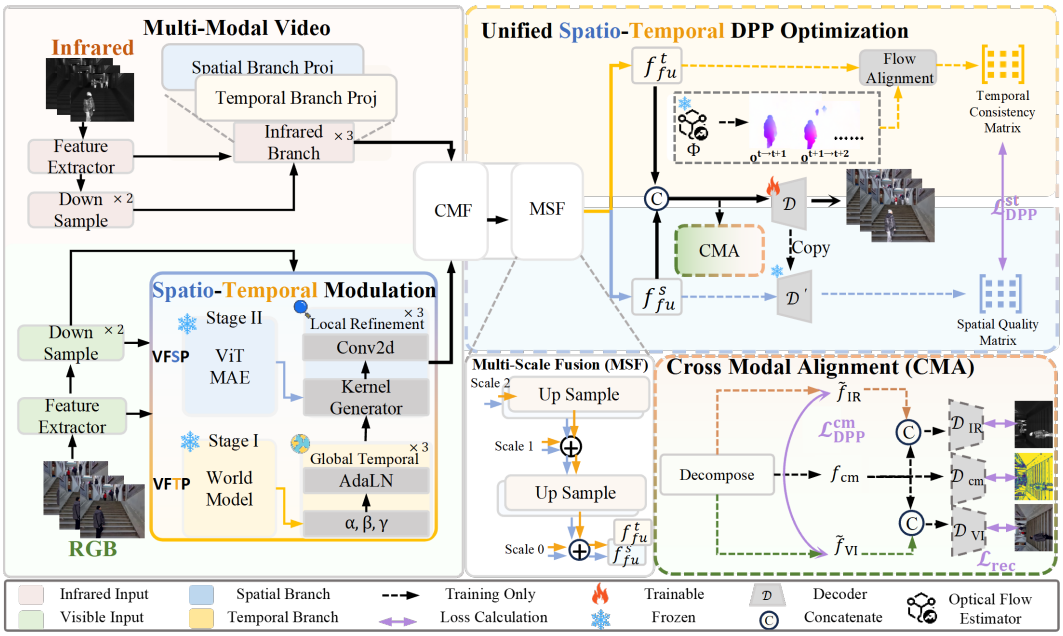

Figure 2: Architecture of D^2Fusion, which comprises spatio-temporal modulation, unified spatio-temporal DPP optimization, and cross-modal alignment. Dashed boxes are used only during training. Cross-Modal Fusion (CMF) combines visible and infrared features via MHCA, while Multi-Scale Fusion (MSF) aggregates multi-scale features to form multi-granularity representations.

2.3 Determinantal Point Process

Determinantal Point Process (DPP) (Kulesza & Taskar, 2012) is a probabilistic model for subset selection that balances diversity and representativeness, having been widely applied to various vision tasks (Yuan & Kitani, 2019; An et al., 2020; Sun et al., 2025; Kazimi et al., 2026). It constructs a kernel matrix $L = \mathbf{QKQ}^T$, where the quality term $\mathbf{Q}$ encodes the individual importance of each item, and the similarity matrix $\mathbf{K}$ measures pairwise resemblance between items together forming a positive semidefinite kernel that enables multi-objective optimization. For a candidate subset s, its probability is proportional to the determinant of the submatrix L_s, which geometrically measures the volume spanned by the selected items. The more diverse the items, the larger the volume $\det(L_s)$ and thus the higher the probability. This inherent mechanism of penalizing redundancy makes DPP particularly suitable for tasks requiring representative subset selection.

3 Methodology

Given a degraded infrared video $V_{\text{IR}} \in \mathbb{R}^{T\times 1\times H\times W}$ and a degraded visible video $V_{\text{VI}} \in \mathbb{R}^{T\times 3\times H\times W}$, with per-frame representations $I_{\text{IR}} \in \mathbb{R}^{1\times H\times W}$ and $I_{\text{VI}} \in \mathbb{R}^{3\times H\times W}$, IVVF aims to produce a fused video with T fused frames $I_{fu} \in \mathbb{R}^{3\times H\times W}$. As shown in Fig. 2, D^2Fusion consists of three main components: spatio-temporal dual-stream modulation, unified spatio-temporal DPP optimization, and structure-aware cross-modal alignment.

3.1 Dual-stream fusion framework

As illustrated in Fig. 3, D^2Fusion adopts a dual-stream architecture with a factorized design, where Stage I focuses on temporal feature learning and Stage II focuses on spatial feature learning. For stage I, the temporal stream is modulated by global temporal modulation weights generated by Video Foundation Temporal Prior (VFTP) f_{VFTP}, which utilizes a world model to encode long-range temporal dependency. Learnable Query-Guided Cross-Attention (LGCA) is constructed to

generate the scaling factor γ, shift parameter β, and gate parameter α as shown in Eq.(1):

$$\{\gamma, \beta, \alpha\} = \text{MLP}\big(\text{Attn}(\mathbf{Q}^t, \mathbf{K}^t, \mathbf{V}^t)\big), \tag{1}$$

where query $\mathbf{Q}^t \in \mathbb{R}^{T\times 1\times C}$ denotes a temporal modulation query, and key $\mathbf{K}^t$, value $\mathbf{V}^t \in \mathbb{R}^{T\times(H\times W)\times C}$ denotes f_{VFTP}. Here, T, H, and W are the numbers of temporal, height, and width tokens. Multi-Head Cross Attention is performed over the spatial axis of f_{VFTP} to pool spatial information into a set of modulation coefficients. Once the coefficients are obtained, Adaptive Layer Normalization (AdaLN) is applied to the original visible features f_{vi} obtained from feature extractor, yielding the temporal-modulated visible feature $\hat{f}^t_{vi}$ as Eq.(2):

$$\hat{f}^t_{vi} = \alpha \cdot \text{Attn}(\hat{\mathbf{Q}}^{\mathbf{t}}, \hat{\mathbf{K}}^{\mathbf{t}}, \hat{\mathbf{V}}^{\mathbf{t}}) + (1-\alpha)\cdot f_{vi}, \tag{2}$$

where $\hat{\mathbf{Q}}^{\mathbf{t}}, \hat{\mathbf{K}}^{\mathbf{t}}, \hat{\mathbf{V}}^{\mathbf{t}}$ are all scaled visible features $\gamma \cdot \frac{f_{vi}-\mu}{\sigma} + \beta$. For stage II, the dual local spatial refinement stream is guided by a Video Foundation Spatial Prior (VFSP) f_{VFSP}, the output features of a ViT-based MAE supervised at the pixel level, producing 2D convolutional kernels K for local spatial structure preservation. Specifically, a group of kernels is generated via another LGCA over the temporal axis. Group convolution Krizhevsky et al. (2017) is adopted for parameter efficiency, with the output reshaped into group-wise kernels. To match patch-level encoder features, the generated kernels are applied as a dilated convolution via sparse sampling at intervals. This yields spatial-modulated visible feature $\hat{f}^s_{vi}$, as formulated in Eq.(3):

$$\hat{f}^s_{vi} = (f \times K)_{i,j} = \sum_{p=0}^{k_h-1}\sum_{q=0}^{k_w-1} K_{p,q}\cdot f_{i+p\cdot\lfloor\frac{H}{h}\rfloor, j+q\cdot\lfloor\frac{W}{w}\rfloor}, \tag{3}$$

where (k_h, k_w) is the convolution 2d kernel size with intervals $(\lfloor\frac{H}{k_h}\rfloor, \lfloor\frac{W}{k_w}\rfloor)$. As shown in Fig. 3, the spatio-temporal modulation is built upon a residual structure where $\hat{f}^t_{vi}$ is detached as the feature map f, fed into the spatial modulation branch, and processed by a convolution $f \times K$ to produce the spatially modulated feature $\hat{f}^s_{vi}$. Such a global-to-local framework integrates global temporal modulation with local spatial refinement, yielding temporally-aware representations while preserving structural fidelity. The hierarchical composition of modulations, together with the stop-gradient on $\hat{f}^t_{vi}$, further enables stable stream-wise optimization. D^2Fusion constructs multi-scale visible features via progressive downsampling, paired with a symmetric infrared branch for cross-modal coordination. Features are decomposed into spatial and temporal streams, which are fused by Cross Modal Fusion (CMF) as Eq.(4):

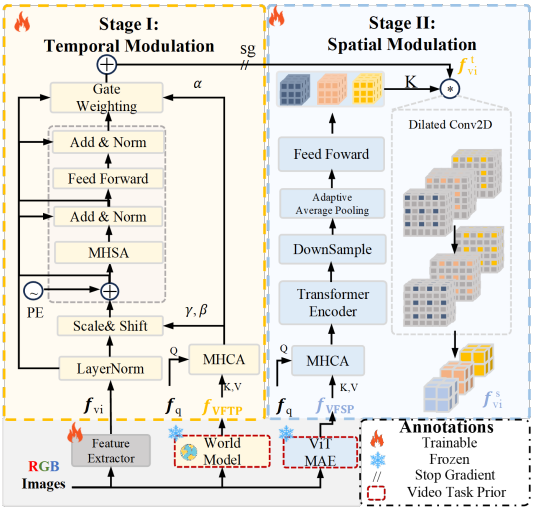

Figure 3: Spatio-Temporal Modulation Module.

$$\begin{cases} \hat{f}^s_{fu} = \text{Attn}(\mathbf{Q}=\hat{f}^s_{vi}, \mathbf{K}, \mathbf{V}=\hat{f}^s_{ir}) + \text{Attn}(\mathbf{Q}=\hat{f}^s_{ir}, \mathbf{K}, \mathbf{V}=\hat{f}^s_{vi}); \\ \hat{f}^t_{fu} = \text{Attn}(\mathbf{Q}=\hat{f}^t_{vi}, \mathbf{K}, \mathbf{V}=\hat{f}^t_{ir}) + \text{Attn}(\mathbf{Q}=\hat{f}^t_{ir}, \mathbf{K}, \mathbf{V}=\hat{f}^t_{vi}), \end{cases} \tag{4}$$

where s, t denotes spatial and temporal branch, respectively. $\hat{f}_{ir}$ and $\hat{f}_{vi}$ are the infrared and spatio-temporal modulated visible features, resulting in the fused feature $\hat{f}_{fu}$. Cross-scale aggregation is then performed via progressive upsampling in Multi-Scale Fusion (MSF), where the temporal stream is further augmented with bi-temporal co-attention (Tang et al., 2026). Iterative fusion from the coarsest scale 2 to the finest scale 0 yields the final outputs f^t_{fu} and f^s_{fu}.

3.2 Unified Spatio-Temporal DPP Optimization

The global temporal modulation and local spatial structure refinement are jointly optimized under a unified DPP objective, which inherently achieves a balance between temporal consistency and

spatial fidelity across dual branches. D^2Fusion regards temporal consistency as the similarity matrix of items in DPP and promotes it by encouraging similarity while suppressing diversity among timestep items. To this end, forward optical flow fields between adjacent frames are computed with a pretrained optical flow estimator Φ during training, yielding $\mathcal{O} = \{o^{1\to2}, ..., o^{T-1\to T}\}$, which is used to define the temporal similarity between items. Specifically, given a target frame $i \in [1, T]$, temporally modulated feature of any preceding frame $j < i$, denoted as $f_{fu}^{t_j}$, is progressively warped along the optical flow chain from j to i. This sequential warping produces aligned temporal feature $f_{fu}^{t_j\to t_i}$, effectively calibrating historical step j to the target step i as in Eq.(5):

$$f_{fu}^{t_j\to t_i} = \mathcal{W}(...\mathcal{W}(\mathcal{W}(f_{fu}^{t_j}, o^{j\to j+1}), o^{j+1\to j+2}), ..., o^{i-1\to i}), \tag{5}$$

where $\mathcal{W}$ represents flow warp operator. The similarity between any pair of warped features $f_{fu}^{t_j\to t_i}$ and $f_{fu}^{t_k\to t_i}$ from frames j and k with respect to frame i is measured by the ℓ_1 Laplacian kernel, yielding the similarity score $t_{j,k}$ as defined in Eq.(6):

$$\mathbf{T}^i = \left[t_{j,k}^i\right]_{i\times i}, \quad t_{j,k}^i = \exp(-\frac{|f_{fu}^{t_j\to t_i} - f_{fu}^{t_k\to t_i}|_1}{\sigma}). \tag{6}$$

This full-chain warping enforces high-order temporal coherence via accumulated optical flow along $j \to i$ path, capturing non-linear motion trajectories beyond frame-to-frame constraints. To address this, a reliability weight s_i is defined for each frame by feeding the spatial-branch feature f_{fu}^s into a frozen replica of the final decoder $\mathcal{D}'$, yielding $\mathcal{F}^s = \mathcal{D}'(f_{fu}^s)$. Since a large spatial standard deviation of $\mathcal{F}^s$ indicates unstable features arising from motion edges, occlusion, or transient illumination, enforcing temporal consistency there risks over-smoothing, which in turn introduces motion ghosting artifacts. We therefore formulate s_i as the negative exponent of the unbiased spatial standard deviation, so that temporal consistency is enforced mainly on frames with stable spatial statistics. The reliability weights up to frame i are organized into a diagonal matrix S^i as Eq.(7):

$$S^i = \text{diag}\{s_0, s_1, \cdots, s_i\}, \quad s_i = \exp\left[-\frac{1}{HW-1}\sqrt{\sum_{u=1}^{H}\sum_{v=1}^{W}(\mathcal{F}_{uv}^s - \bar{\mathcal{F}}^s)^2}\right], \quad \mathcal{F}^s = \mathcal{D}'(f_{fu}^s). \tag{7}$$

S^i serves as a reliability term, together with $\mathbf{T}^i$ forming the spatio-temporal DPP kernel $S^i\mathbf{T}^i(S^i)'$. On one hand, the temporal similarity matrix $\mathbf{T}$ encodes pairwise feature affinities across frames, where minimizing its determinant is equivalent to enforcing linear dependence among temporal feature vectors, thereby penalizing abrupt variations. On the other hand, the spatial reliability matrix S down-weights frames with unstable spatial statistics to prevent temporal over-smoothing. This multiplicative coupling jointly encourages spatio-temporal consistency. Accordingly, the positive semi-definite spatio-temporal kernel $S^i\mathbf{T}^i(S^i)'$ is constructed. Its determinant is minimized via the unified spatio-temporal DPP loss $\mathcal{L}_{DPP}^{st}$, as in Eq.(8):

$$\mathcal{L}_{DPP}^{st} = \sum_{i=1}^{t} \log\det(S^i\mathbf{T}^i(S^i)' + \epsilon \cdot \mathbb{I}), \tag{8}$$

where $\epsilon = 10^{-6}$ is a small positive constant and $\mathbb{I}$ is the identity matrix for Tikhonov regularization (Golub et al., 1999) to avoid singularity.

3.3 Structure-Aware Cross-Modal Alignment via Optical Flow Distillation

Existing methods neglect the inherent geometric misalignment between infrared and visible domains, causing thermal targets to lose spatial correspondence with visible structures. To address this, we introduce a geometric distillation paradigm that explicitly models cross-modal spatial correspondence. During training, a pretrained optical flow estimator provides a dense spatial offset field from infrared–visible pairs as modality-invariant structural guidance, based on which we reformulate the reconstruction loss into a structural fidelity objective. This objective distills a pixel map $\mathbf{M}_s$ that captures modality-invariant positional and structural cues while disentangling them from modality-specific textures. $\mathbf{M}_s$ is constructed via optical-flow alignment, gradient pooling, and flat-mask-guided label smoothing (see Alg. 3 in Appendix for details).

A triple-branch feature decomposition module is then designed to disentangle features into cross-modal shared f_{cm}, infrared-specific $\tilde{f}_{ir}$, and visible-specific $\tilde{f}_{vi}$ representations. To ensure that the modality-specific representations encode genuinely unique and non-redundant information, reconstruction constraints and diversity regularization are employed. For reconstruction constrains $\mathcal{L}_{rec}$, the cross-modal shared representation f_{cm} is trained to reconstruct the modality-invariant structural map $\mathbf{M}_s$, while f_{cm} combined with each modality-specific code $\tilde{f}_{ir/vi}$ reconstructs the corresponding modality input $I_{\text{IR/VI}}$. Each modality is decoded by an auxiliary decoder $\mathcal{D}_{\text{IR/VI}}$, thereby capturing the common geometry across modalities, as formulated in Eq.(9):

$$\mathcal{L}_{rec}^{cm} = \|\mathcal{D}_{cm}(f_{cm}) - \mathbf{M}_s\|_2; \qquad \mathcal{L}_{rec}^{\text{IR/VI}} = \|\mathcal{D}_{\text{IR/VI}}(f_{cm} \copyright \tilde{f}_{\text{IR/VI}}) - I_{\text{IR/VI}}\|_2; \quad \mathcal{L}_{rec} = \mathcal{L}_{rec}^{cm} + \mathcal{L}_{rec}^{\text{IR/VI}}, \tag{9}$$

where © denotes concatenation. However, if the modality-specific components of the two modalities remain highly similar, they become largely redundant and the fusion gain diminishes. Explicitly reducing their similarity thus promotes inter-modal diversity, encouraging each modality to retain discriminative cues unavailable from the other. To enrich cross-modal representation, we introduce a diversity regularization that encourages $\tilde{f}_{ir}$ and $\tilde{f}_{vi}$ to be mutually exclusive. Specifically, non-overlapping patches are unfolded into patch-level tokens $\tilde{f}_{ir}^{(p)}$ and $\tilde{f}_{vi}^{(p)}$ for the p-th patch. A modality-specific cosine similarity matrix $\mathbf{K}^p$ is then constructed over the two modalities. $\mathbf{K}^p$ quantifies per-patch redundancy between modalities and serves as a diversity constraint in Eq.(10):

$$\mathbf{K}^p = \begin{pmatrix} \langle \tilde{f}_{ir}^{(p)}, \tilde{f}_{ir}^{(p)} \rangle & \langle \tilde{f}_{ir}^{(p)}, \tilde{f}_{vi}^{(p)} \rangle \\ \langle \tilde{f}_{vi}^{(p)}, \tilde{f}_{ir}^{(p)} \rangle & \langle \tilde{f}_{vi}^{(p)}, \tilde{f}_{vi}^{(p)} \rangle \end{pmatrix} = \begin{pmatrix} 1 & cos\theta_p \\ cos\theta_p & 1 \end{pmatrix}, \tag{10}$$

where $\langle \cdot, \cdot \rangle$ denotes the calculation of cosine similarity $cos\theta_p$ between different modalities in the p-th patch. While the diversity term captures modality-specific content, the quality term is designed to measure how well the fused image reconstructs the structure shared across modalities. Since gradients reflect structural content, we measure reconstruction quality by the gradient consistency between the fused image and the multi-modal reference. A higher reconstruction quality indicates that the underlying shared structure is faithfully preserved, and therefore encourages stronger preservation of modality-specific content. We therefore introduce a per-patch quality score q^p that measures the gradient consistency between the fused image I_{fu} and the multi-modal reference $I_{\text{IR/VI}}$ in the p-th patch Ω^p. Specifically, gradient magnitude maps G_{fu}, G_{IR} and G_{VI} are computed on the Y channel of YC_bC_r, with per-pixel differences $\Delta G = |G_{fu} - G_{\text{IR}}|_1 + |G_{fu} - G_{\text{VI}}|_1$. Averaging $\Delta G(u, v)$ over Ω^p and mapping the average through a negative exponential yields q^p, which forms the diagonal quality weight matrix $\mathbf{Q}^p$ as in Eq.(11):

$$\mathbf{Q}^p = \begin{pmatrix} q^p & 0 \\ 0 & q^p \end{pmatrix}, \quad q^p = \exp\left(-\frac{1}{|\Omega^p|} \sum_{(u,v)\in\Omega^p} \Delta G(u, v) \right), \tag{11}$$

where q^p weights cross-modal kernel matrix per patch, so that poorly reconstructed patches with low q^p contribute less to diversity objective. The patch-wise DPP kernel is $\mathbf{Q}^p\mathbf{K}^p(\mathbf{Q}^p)'$, and its determinant is maximized via cross-modal loss $\mathcal{L}_{DPP}^{cm}$ over all N patches, as formulated in Eq.(12):

$$\mathcal{L}_{DPP}^{cm} = -\frac{1}{N} \sum_{p=1}^{N} \log\det(\mathbf{Q}^p\mathbf{K}^p(\mathbf{Q}^p)' + \epsilon \cdot \mathbb{I}). \tag{12}$$

4 Experiments

4.1 Experimental Settings

Datasets. We train on 200 videos from M3SVD (Tang et al., 2026) and test on the remaining 20, including 6 nighttime sequences that constitute degraded scenarios. For cross-dataset validation, we directly test on all 24 scenarios of HDO (Xie et al., 2024a), which represent normal scenarios. M3SVD is a large-scale infrared-visible video benchmark with 220 aligned pairs and $153,979$ frames covering low-light, overexposure, camouflage, and occlusion, while HDO contains $5,788$ aligned pairs at 480×640, characterized by fast-moving objects.

Implementation Details. We set the time step $T = 6$ for training and inference. We use VideoMAE as VFTP and V-JEPA 2.0 as VFSP; the generated kernel size (h, w) is configured as $(8, 8)$, and the

Table 1: Quantitative comparison on the M3SVD Tang et al. (2026) dataset. Each video in M3SVD contains 200 frames. The best results are highlighted in bold. '$*$' represents IVVF methods.

Methods	Multi-Modal Multi-Scene Video Dataset(M3SVD)					
	EN ↑	MI ↑	SD ↑	SSIM ↑	flowD ↓	feaCD ↓
U2Fusion (Xu et al., 2020)	6.904	2.490	35.731	0.600	6.547	-
LRRNet (Li et al., 2023)	6.889	3.120	38.251	0.609	6.874	-
MURF (Xu et al., 2023)	6.185	1.825	22.731	0.543	6.840	-
DDFM (Zhao et al., 2023)	6.750	2.656	32.000	0.609	6.778	-
TC-MOA (Zhu et al., 2024)	7.095	2.800	42.412	0.593	5.102	-
TIMFusion (Liu et al., 2024c)	7.063	3.015	50.824	0.580	5.890	-
GIFNet (Cheng et al., 2025)	7.042	2.548	44.207	0.580	6.294	-
RCVS (Xie et al., 2024a)	6.551	1.979	37.468	0.585	6.443	-
Temcoco* (Gong et al., 2025)	**7.169**	3.529	50.678	0.593	4.378	1.890
VideoFusion* (Tang et al., 2026)	7.163	4.019	**52.654**	0.632	3.556	1.904
D^2Fusion *	7.110	**4.126**	51.601	**0.635**	**3.343**	**1.812**

temperature coefficient in constructing the temporal consistency matrix is set to $\sigma = 10$. The patch size for quality score computation in cross-modal spatial alignment is $k = 16$.

Compared Methods. We compare D^2Fusion against SOTA methods in both IVIF and IVVF, covering CNN-based (Xu et al., 2020; Li et al., 2023; Xu et al., 2023; Zhu et al., 2024; Liu et al., 2024c; Cheng et al., 2025), Transformer-based (Gong et al., 2025; Tang et al., 2026), and diffusion-based architectures (Zhao et al., 2023). To ensure a fair comparison, we follow Gong et al. (2025); Tang et al. (2026) by applying DSTNet (Pan et al., 2023) for deblurring visible frames and MDIVD-net (Cai et al., 2024) for destriping infrared frames when evaluating IVIF methods.

Evaluation Metrics. For quantitative evaluation, we adopt four spatial quality metrics: EN (Roberts et al., 2008), MI (Qu et al., 2002), SD (Rao, 1997), SSIM (Wang et al., 2004), and two temporal consistency metrics: flowD (Gong et al., 2025) and feaCD (Gong et al., 2025), which measure consistency at the pixel and feature levels respectively. To prevent over-smoothing induced by temporal artifacts in individual frames, we further introduce Edge Intensity (EI) (Wang et al., 2025), which is particularly emphasized in the HDO cross-domain validation given its high-speed motion scenes.

4.2 Comparison with State-of-the-Arts

Quantitative Comparisons. Tab. 1 presents quantitative comparisons among competing methods in degraded M3SVD scenarios. D^2Fusion excels in spatial fidelity with the highest MI and SSIM, demonstrating its ability to maximally exploit the complementary spatial information offered by two modalities. D^2Fusion also achieves the best flowD and feaCD, indicating excellent temporal consistency across both feature and pixel levels. Among IVIF methods, flowD is suboptimal as they overlook temporal context. Moreover, cross-dataset validation is performed on HDO to verify cross-scene generalization. Tab. 2 shows D^2Fusion gets the highest EN, SSIM, which indicates its superior spatial preservation capacity. For temporal consistency, D^2Fusion achieves the best feaCD and second-best flowD. Although VideoFusion attains a relatively low flowD, its inferior feaCD and EI reveal that its temporal smoothness stems from over-smoothing and cross-time-step motion leakage rather than meaningful temporal modeling, whereas our method achieves superior feaCD in feature level.

Table 2: Cross dataset validation on HDO dataset.

Methods	High-quality Dual-Optical Dataset (HDO)				
	EN↑	SSIM↑	EI↑	flowD↓	feaCD↓
LRRNet	6.647	0.584	43.438	7.305	-
GIFNet	6.869	0.623	45.218	6.585	-
TIMFusion	6.663	0.589	36.292	7.274	-
TemCoCo*	6.597	0.591	49.862	6.747	1.869
VideoFusion*	6.980	0.621	50.940	**4.669**	1.872
D^2Fusion*	**7.017**	**0.626**	**57.324**	6.014	**1.776**

Qualitative Comparisons. Fig. 4 (I)-(II) shows IVIF methods fail to preserve salient targets, whereas IVVF baselines and D^2Fusion successfully capture them with broader highlighted fusion regions under degraded scenarios. However, TemCoCo introduces color bleeding artifacts on salient



Figure 4: Qualitative comparison on M3SVD Tang et al. (2026) and HDO Xie et al. (2024a). Salient target regions and background detail regions that require emphasis are amplified.

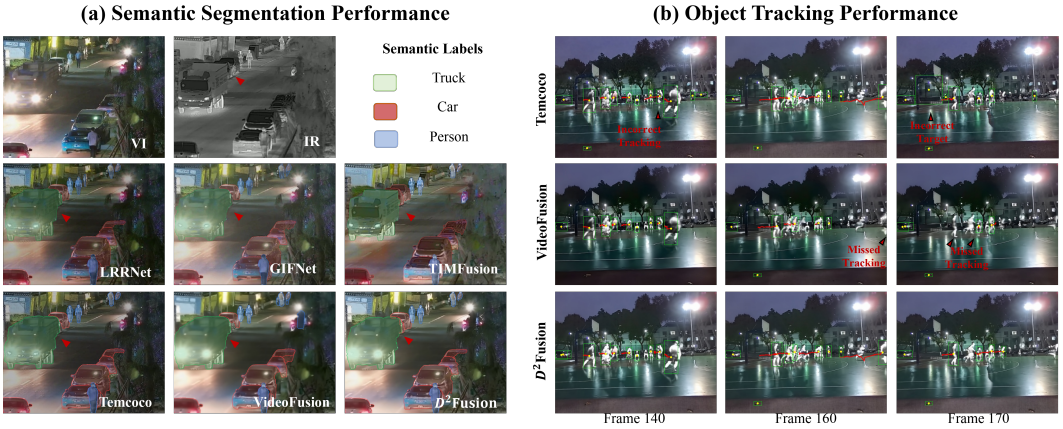

Figure 5: Downstream task visualization. (a) D^2Fusion correctly segments the occluded red-arrow car. (b) D^2Fusion tracks correctly and completely.

targets in nighttime scenes, and VideoFusion causes excessive temporal smoothing at the cost of severe motion artifacts. In contrast, D^2Fusion strikes a principled balance between spatial structures and temporal coherence. In Fig. 4 (III), we also test cross dataset validation on normal scenario. D^2Fusion renders the motorcyclist (red box) with the sharpest contours and the most salient target presentation, while the moving car (green box) is free from noticeable motion artifacts.

Extension Experiments. IVVF can serve as input to multiple downstream vision tasks. We feed the fused videos from all methods into semantic segmentation and object tracking tasks. In semantic segmentation, as shown in Fig. 5(a), the car indicated by the red arrow is occluded by a truck with illuminated lights, appearing only in the infrared modality while remaining invisible in the visible one. Only TIMFusion and D^2Fusion segment this car successfully. Unlike TIMFusion, which is driven by a downstream segmentation task, our method requires no task-specific supervision yet still preserves sufficient semantic information for accurate segmentation. For object tracking, Fig. 5(b) compares IVVF baselines and D^2Fusion, all using YOLO v11 as the tracker. In nighttime sports

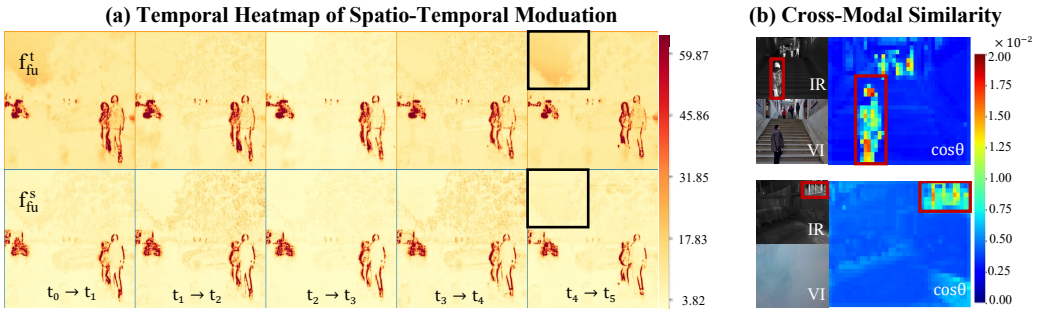

Figure 6: Visualization of features. (a) Per-pixel temporal variance of f^t_{fu} and f^s_{fu} in spatio-temporal modulation. (b) Patch-wise cross-modal similarity for modality-specific diversity.

scenes, existing baselines exhibit three typical failure modes, namely wrong target, missed tracking, and wrong tracking, whereas D^2Fusion is free from all of them and maintains correct tracking.

4.3 Ablation Study

Effectiveness of Components. To validate the effectiveness of VFP-guided spatio-temporal modulation, unified spatio-temporal DPP optimization, and cross-modal alignment, we conduct ablation studies by removing each component individually. Tab. 3 reports the results on the degraded M3SVD test set. Comparing (1) and (2), removing VFP degrades both spatial and temporal metrics simultaneously, indicating that VFP is essential for preserving spatio-temporal quality. Comparing (3) and (5), the model without the unified spatio-temporal DPP loss achieves competitive spatial metrics such as MI and EI, yet its temporal metric flowD degrades by 2.8%. This suggests that D^2Fusion strikes a better trade-off between spatial and temporal consistency and reaches a superior Pareto frontier. Finally, comparing (4) and (5), removing cross-modal alignment degrades spatial metrics such as MI and EI, due to the resulting lack of cross-modal spatial correspondence.

Table 3: Ablation studies.

	VFP	$\mathcal{L}^{st}_{DPP}$	$\mathcal{L}^{cm}_{DPP}$	MI↑	EI↑	flowD↓	feaCD↓
(1)				3.989	52.805	3.506	1.822
(2)	✓			4.107	53.094	3.409	1.817
(3)	✓		✓	4.120	53.666	3.436	1.811
(4)	✓	✓		4.112	52.796	3.388	1.814
(5)	✓	✓	✓	4.126	53.510	3.343	1.812

Activation Heatmaps over Time. Fig. 6(a) visualizes per-pixel temporal variance at two stages. Temporally modulated f^t_{fu} extracts inter-frame motion cues but also mixes in background fluctuations from noise and illumination jitter. Conditioned on f^t_{fu}, the spatial modulation f^s_{fu} suppresses these redundant perturbations on static backgrounds while preserving motion responses on moving objects, validating that the spatio-temporal decomposition separates motion from temporal jitter.

Cross-Modal Similarity Analysis. Fig. 6(b) visualizes patch-wise infrared-visible similarity $cos\theta_p$. Heat-source regions exhibit higher similarity, as both modalities describe the same salient structure there, whereas background statistics remain largely modality-specific. Across all patches, similarity stays below 0.02, indicating that modality-specific information is preserved rather than collapsed.

5 Conclusion

In this paper, we address spatio-temporal optimization imbalance and insufficient cross-modal exploitation in infrared and visible video fusion. We propose D^2Fusion, a decomposed paradigm that integrates temporal-then-spatial branches, using Video Foundation Priors for global temporal modulation and local spatial refinement. By regarding temporal consistency and spatial richness as a quality–diversity trade-off, the unified DPP objective balances them adaptively, while a modality-invariant prior over shared geometric structure counters cross-modal misalignment. Experiments show that D^2Fusion achieves state-of-the-art performance on M3SVD with strong cross-dataset generalization, facilitating the practical application of infrared and visible video fusion.

AI USE STATEMENT

In this work, we used generative AI tools for none of the tasks requiring disclosure. We have not used generative AI tools for any other tasks requiring disclosure, and all remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for none of the tasks with recommended disclosure. We have reviewed all work in this paper, which was carried out entirely by the authors without AI assistance. We take responsibility for the final content of this work, including text, claims, or artifacts produced without the aid of generative AI.

REPRODUCIBILITY STATEMENT

We aim to make all results reported in the paper reproducible. Datasets, evaluation metrics, and hyper-parameters in our experiments are described in 4.1. Since D^2Fusion utilizes large-scale pre-trained video models, V-JEPA 2.0, and VideoMAE as VFTP and VFSP, we described their details in A.2. Moreover, the supplementary material includes a .zip file containing an anonymized repository with training and evaluation scripts, configuration files, hyper-parameters, model variants, seeds, and environment specifications.

REFERENCES

Bang An, Jie Lyu, Zhenyi Wang, Chunyuan Li, Changwei Hu, Fei Tan, Ruiyi Zhang, Yifan Hu, and Changyou Chen. Repulsive attention: Rethinking multi-head attention as bayesian inference. In *EMNLP*, pp. 236–255, 2020.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Lijing Cai, Xiangyu Dong, Kailai Zhou, and Xun Cao. Exploring video denoising in thermal infrared imaging: Physics-inspired noise generator, dataset, and model. *IEEE Transactions on Image Processing*, 33:3839–3854, 2024.

Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Zeyang Zhang, Sara Atito, Muhammad Awais, and Josef Kittler. One model for all: Low-level task interaction is a key to task-agnostic image fusion. In *CVPR*, pp. 28102–28112, 2025.

Jianfeng Ding, Hao Zhang, Zhongyuan Wang, Jinsheng Xiao, Xin Tian, Zhen Han, and Jiayi Ma. Cmvf: Cross-modal unregistered video fusion via spatio-temporal consistency. *Information Fusion*, pp. 104212, 2026.

Gene H Golub, Per Christian Hansen, and Dianne P O'Leary. Tikhonov regularization and total least squares. *SIAM journal on matrix analysis and applications*, pp. 185–194, 1999.

Meiqi Gong, Hao Zhang, Xunpeng Yi, Linfeng Tang, and Jiayi Ma. Temcoco: Temporally consistent multi-modal video fusion with visual-semantic collaboration. In *ICCV*, pp. 14326–14335, 2025.

Yuqi Han, Zhihui Zheng, Weijian Su, Mingkai Wei, Liang Zhang, Jinli Suo, and Qiang Zhang. Seq-if: Sequentially consistent infrared-visible video fusion under time-varying illumination for perception enhancement. *IEEE TCSVT*, 2026.

Yongchao Huang. Vjepa: Variational joint embedding predictive architectures as probabilistic world models. *arXiv preprint arXiv:2601.14354*, 2026.

Tahira Kazimi, Connor Dunlop, and Pinar Yanardag. Diverse video generation with determinantal point process-guided policy optimization. In *CVPR*, pp. 12839–12848, 2026.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, pp. 123–286, 2012.

Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE TIP*, pp. 2614–2623, 2018.

Hui Li and Xiao-Jun Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, pp. 102147, 2024.

Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE TPAMI*, pp. 11040–11052, 2023.

Xilai Li, Weijun Jiang, Xiaosong Li, Yang Liu, Hongbin Wang, Tao Ye, Huafeng Li, and Haishu Tan. Mavfusion: Efficient infrared and visible video fusion via motion-aware sparse interaction. *arXiv preprint arXiv:2604.01958*, 2026a.

Xingyuan Li, Haoyuan Xu, Shulin Li, Xiang Chen, Zhiying Jiang, and Jinyuan Liu. Drfusion: Drift-resilient temporally consistent infrared-visible video fusion. *arXiv preprint arXiv:2605.25775*, 2026b.

Jinyuan Liu, Runjia Lin, Guanyao Wu, Risheng Liu, Zhongxuan Luo, and Xin Fan. Coconet: Coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. *IJCV*, pp. 1748–1775, 2024a.

Jinyuan Liu, Guanyao Wu, Zhu Liu, Di Wang, Zhiying Jiang, Long Ma, Wei Zhong, Xin Fan, and Risheng Liu. Infrared and visible image fusion: From data compatibility to task adaption. *IEEE TPAMI*, pp. 2349–2369, 2024b.

Risheng Liu, Zhu Liu, Jinyuan Liu, Xin Fan, and Zhongxuan Luo. A task-guided, implicitly-searched and meta-initialized deep model for image fusion. *IEEE TPAMI*, pp. 6594–6609, 2024c.

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

Edoardo Palladin, Roland Dietze, Praveen Narayanan, Mario Bijelic, and Felix Heide. Samfusion: Sensor-adaptive multimodal fusion for 3d object detection in adverse weather. In *ECCV*, pp. 484–503, 2024.

Jinshan Pan, Boming Xu, Jiangxin Dong, Jianjun Ge, and Jinhui Tang. Deep discriminative spatial and temporal network for efficient video deblurring. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22191–22200. IEEE, 2023.

Jinye Peng, Yu Chen, Jun Wang, Shenglin Peng, Shuyi Qu, and Zhaoke Liu. Drfusion: Detail-retention framework for infrared and visible image fusion in complex illumination scenes. *Infrared Physics & Technology*, pp. 106038, 2025.

Zhenkai Qiang, Yu Shen, Guanxiang Pei, and Zhen Yuan. Ofvfusion: Optical flow-guided saliency learning for infrared and visible video fusion. *Pattern Recognition*, pp. 114678, 2026.

Guihong Qu, Dali Zhang, and Pingfan Yan. Information measure for performance of image fusion. *Electronics letters*, pp. 1, 2002.

Yun-Jiang Rao. In-fibre bragg grating sensors. *Measurement science and technology*, pp. 355–375, 1997.

J Wesley Roberts, Jan A Van Aardt, and Fethi Babikker Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, pp. 023522, 2008.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp 3: a training-free approach for list-wise frame selection in video-llms. In *ICCV*, pp. 24090–24101, 2025.

Linfeng Tang, Yeda Wang, Meiqi Gong, Zizhuo Li, Yuxin Deng, Xunpeng Yi, Chunyu Li, Han Xu, Hao Zhang, and Jiayi Ma. Videofusion: a spatio-temporal collaborative network for multi-modal video fusion. In *CVPR*, pp. 19559–19569, 2026.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, pp. 402–419, 2020.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, pp. 10078–10093, 2022.

Rongchao Wang, Zhaofa Zhou, Shuhui Li, and Zhili Zhang. Advances and challenges in infrared-visible image fusion: A comprehensive review of techniques and applications. *Artificial Intelligence Review*, pp. 18, 2025.

Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. In *ECCV*, pp. 36–54, 2024.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, pp. 600–612, 2004.

Guanyao Wu, Haoyu Liu, Hongming Fu, Yichuan Peng, Jinyuan Liu, Xin Fan, and Risheng Liu. Every sam drop counts: Embracing semantic priors for multi-modality image fusion and beyond. In *CVPR*, pp. 17882–17891, 2025.

Housheng Xie, Meng Sang, Yukuan Zhang, Yang Yang, Shan Zhao, and Jianbo Zhong. Rcvs: A unified registration and fusion framework for video streams. *IEEE Transactions on Multimedia*, pp. 11031–11043, 2024a.

Yaofeng Xie, Lingwei Kong, Kai Chen, Ziqiang Zheng, Xiao Yu, Zhibin Yu, and Bing Zheng. Uveb: A large-scale benchmark and baseline towards real-world underwater video enhancement. In *CVPR*, pp. 22358–22367, 2024b.

Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE TPAMI*, pp. 502–518, 2020.

Han Xu, Jiteng Yuan, and Jiayi Ma. Murf: Mutually reinforcing multi-modal image registration and fusion. *IEEE TPAMI*, pp. 12148–12166, 2023.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, and Dacheng Tao. Gmflow: Learning optical flow via global matching. In *CVPR*, pp. 8111–8120, 2022.

Ye Yuan and Kris Kitani. Diverse trajectory forecasting with determinantal point processes. *arXiv preprint arXiv:1907.04967*, 2019.

Xiaodong Zhang, Qing Hu, Kun Zhang, Minghao Jiang, and Shaoshu Gao. Smoothfusion: Smooth-transition adaptive keyframe registration and perceptual fusion framework for infrared-visible video streams. *Optics & Laser Technology*, pp. 115692, 2026.

Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: Denoising diffusion model for multi-modality image fusion. In *ICCV*, pp. 8048–8059, 2023.

Zixiang Zhao, Haowen Bai, Bingxin Ke, Yukun Cui, Lilun Deng, Yulun Zhang, Kai Zhang, and Konrad Schindler. A unified solution to video fusion: From multi-frame learning to benchmarking. In *NeurIPS*, pp. 128504–128535, 2026.

Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *CVPR*, pp. 7099–7108, 2024.

A APPENDIX

A.1 IMPLEMENTATION DETAILS

A.1.1 FLOW VALID MASK GENERATION

Since training uses cropping as a preprocessing step, the optical flow $\mathcal{O}$ computed between adjacent frames must be cropped to the same region. This requires a flow valid mask $\mathbf{M}_v$, whose generation is detailed in Alg. 1. Given the cropped top border h_0, left border w_0 and patch size P, we first load pretrained SEA-RAFT (Wang et al., 2024) as an optical flow estimator Φ to calculate adjacent-frame flow $\mathcal{O} = \{o^{1\to2}, ..., o^{T-1\to T}\} \in \mathbb{R}^{(T-1)\times H\times W}$ and then crop the target $P \times P$ region $\{(h_0, w_0), (h_0, w_0 + P), (h_0 + P, w_0 + P), (h_0 + P, w_0)\}$. Then we compute a binary valid mask per pixel. We initialize an all-one mask $\mathbf{M}_v^1$ for the first time step, and sequentially warp the mask at step $t-1$ with the estimated flow $o^{t-1\to t}$. The mask at step t, $\mathbf{M}_v^t$ is obtained by logically ANDing the warped mask at t step with the mask $\mathbf{M}_v^{t-1}$ at the last step $t-1$. Finally, we attain estimated optical flow $\mathcal{O}$ between adjacent frames and its corresponding mask $\mathbf{M}_v$ for the following calculation of temporal consistency matrix $\mathbf{T}$.

Algorithm 1 Flow Valid Mask Generation

Input: Visible sequence $V_{\text{VI}} \in \mathbb{R}^{T\times C\times H\times W}$, patch size P, top-left coordinates (h_0, w_0), optical flow estimator Φ.
Output: Cropped flow $\hat{\mathcal{O}} \in \mathbb{R}^{(T-1)\times P\times P}$, flow valid mask $\mathbf{M}_v \in \{0,1\}^{T\times P\times P}$.
Step 1: Flow estimator initialization
  $\Phi \leftarrow$ LoadPretrained(). {Load optical flow estimator.}
  Φ.requires_grad_(False). {Freeze parameters.}
Step 2: Adjacent-frame flow estimation
  $\mathcal{O} \leftarrow \Phi(V_{\text{VI}}^{t-1}, V_{\text{VI}}^{t})$. {Estimate flow between previous and current frames.}
Step 3: Patch-wise flow cropping
  $(\Delta h, \Delta w) \leftarrow \text{meshgrid}(\text{arange}(P), \text{arange}(P))$. {Build patch grid.}
  $h \leftarrow h_0 + \Delta h, \quad w \leftarrow w_0 + \Delta w$. {Compute crop coordinates.}
  $\hat{\mathcal{O}} \leftarrow \text{Crop}(\mathcal{O}, h, w)$. {Crop flow to patch.}
Step 4: Flow valid mask initialization
  $\mathbf{M}_v^1 \leftarrow \mathbb{1} \in \mathbb{R}^{P\times P}$. {Initialize all-one mask.}
Step 5: Temporal validity propagation
  for $t = 2$ to T **do**
    $\mathbf{M}_v^t \leftarrow \text{Warp}(\mathbf{M}_v^{t-1}, o^{t-1\to t})$. {Warp previous mask with flow.}
    $\mathbf{M}_v^t \leftarrow \mathbf{M}_v^t \wedge \mathbf{M}_v^{t-1}$. {Logical AND.}
  end for
  $\mathbf{M}_v \leftarrow [\mathbf{M}_v^1, \ldots, \mathbf{M}_v^T]$. {Concatenate masks.}
Step 6: Return
  return $\hat{\mathcal{O}}, M_v$.

Once we get the estimated optical flow $\mathcal{O}$ and its valid mask $\mathbf{M}_v$, we construct the temporal similarity matrix as Alg. 2. For each target frame i, we first align every preceding frame $j < i$ to frame i by progressively warping $\mathcal{W}$ its feature $f_{fu}^{t_j}$ along the optical flow chain $o^{j\to j+1}, ..., o^{i-1\to i}$, yielding the aligned feature $f_{fu}^{t_j\to t_i}$. We then compute the pairwise ℓ_1 difference between every pair of aligned features $f_{fu}^{t_j\to t_i}$ and $f_{fu}^{t_k\to t_i}$, and mask out invalid regions using the corresponding flow valid masks $\mathbf{M}_v$. The resulting difference is passed through a Laplacian kernel to obtain the similarity score $t_{j,k}^i = \exp(-\Delta_{j,k}^i/\sigma)$, which forms the $i \times i$ similarity matrix $\mathbf{T}^i$. Finally, the matrices for all target frames are collected into $\mathbf{T} = \{\mathbf{T}^i\}_{i=1}^T$. $\mathbf{T}$ is the temporal consistency matrix in unified spatio-temporal DPP optimization.

Algorithm 2 Temporal Optical Flow Consistency

Input: Optical flow set $\mathcal{O} = \{o^{1\to 2}, \ldots, o^{T-1\to T}\}$, flow valid mask $\mathbf{M}_v \in \{0,1\}^{T\times P\times P}$, temporal-modulated feature sequence $f_{fu} = \{f_{fu}^{t_1}, \ldots, f_{fu}^{t_T}\}$, temperature σ.
Output: Temporal similarity matrix $\mathbf{T}$.
Step 1: Full-chain flow warping
  for $i = 1$ to T **do**
    for $j = 1$ to i **do**
      if $j = i$ **then**
        $f_{fu}^{t_j\to t_i} \leftarrow f_{fu}^{t_i}$. {Identity, no warping.}
      else
        $f_{fu}^{t_j\to t_i} \leftarrow \mathcal{W}(\cdots \mathcal{W}(\mathcal{W}(f_{fu}^{t_j}, o^{j\to j+1}), o^{j+1\to j+2}), \ldots, o^{i-1\to i})$. {Eq.(5).}
      end if
    end for
  end for
Step 2: Temporal similarity matrix construction
  for $i = 1$ to T **do**
    Collect all aligned features $\{f_{fu}^{t_j\to t_i}\}_{j<i}$ w.r.t. frame i. {Fix target frame i.}
    for $j = 1$ to i **do**
      for $k = 1$ to i **do**
        $\Delta_{j,k}^{i} \leftarrow |f_{fu}^{t_j\to t_i} - f_{fu}^{t_k\to t_i}|_1$. {Pairwise ℓ_1 difference.}
        $\mathbf{M}_{\text{pair}} \leftarrow \mathbf{M}_v^{j\to i} \wedge \mathbf{M}_v^{k\to i}$. {Pairwise valid mask.}
        $\Delta_{j,k}^{i} \leftarrow \Delta_{j,k}^{i} \odot \mathbf{M}_{\text{pair}}$. {Mask invalid pairs.}
        $t_{j,k}^{i} \leftarrow \exp(-\Delta_{j,k}^{i}/\sigma)$. {Eq.(6).}
      end for
    end for
    $\mathbf{T}^i \leftarrow [t_{j,k}^{i}]_{i\times i}$. {Build similarity matrix.}
  end for
  $\mathbf{T} \leftarrow \{\mathbf{T}^i\}_{i=1}^{T}$. {Collect all similarity matrices.}
Step 3: Return
  return $\mathbf{T}$.

A.1.2 Cross-Modal Alignment

To generate the modality-invariant structural mask $\mathbf{M}_s$ in structure-aware cross-modal alignment, the infrared modality frame I_{IR} is aligned to the visible modality I_{VI} using a cross-modal optical flow estimated between the visible and infrared modalities. As shown in Alg. 3, the gradients of each modality, $G_{ir/vi}$, are then computed and average-pooled to suppress random noise. To identify regions lacking significant structure, a flat mask $\mathbf{M}_f$ is constructed via the logical conjunction of thresholded gradient magnitudes from both modalities, using a threshold θ. When the ℓ_1 difference ΔG between the smoothed infrared and visible images is sufficiently small, the two modalities are considered structurally consistent in these non-flat regions. Thus, label smoothing is applied by assigning a soft label $\exp\{-\Delta G/\tau\}$ to these regions. In other words, although these regions are classified as non-flat by $\mathbf{M}_f$, label smoothing relaxes their hard validity into a soft one, reflecting their partial structural consistency across modalities. Finally, Gaussian smoothing is applied to suppress fringe artifacts and produce the refined modality-invariant pixel map $\mathbf{M}_s$.

A.2 Video Foundation Prior

A.2.1 VFTP: V-JEPA 2.0

V-JEPA 2.0 (Huang, 2026) is a self-supervised video model built on the Joint Embedding Predictive Architecture (JEPA) and pretrained on over one million hours of internet video. Its goal is to learn general-purpose representations of the physical world rather than perform pixel-level reconstruction. It is trained by predicting the latent representations of masked video tokens, which forces the encoder to capture predictable spatio-temporal dynamics while discarding unpredictable detail and noise.

Algorithm 3 Modality-Invariant Structure Mask Generation via Cross-Modal Geometric Distillation

Input: Infrared frame $I_{\text{IR}} \in \mathbb{R}^{1\times H\times W}$, visible frame $I_{\text{VI}} \in \mathbb{R}^{3\times H\times W}$, pretrained optical flow estimator Φ, gradient tolerance τ, quantile threshold τ_q.
Output: modality-invariant structure mask $\mathbf{M}_{\text{s}} \in [0,1]^{H\times W}$.
Step 1: Cross-modal alignment
  $\mathcal{F} \leftarrow \Phi(I_{\text{IR}}, I_{\text{VI}})$. {Estimate optical flow $\mathcal{F}$ cross modalities.}
  $\tilde{I}_{ir} \leftarrow \text{Warp}(I_{\text{IR}}, \mathcal{F})$. {Align infrared to visible coordinate space.}
Step 2: Gradient computation and smoothing
  $G_{ir/vi} \leftarrow |\nabla_x \tilde{I}_{\text{IR/VI}}| + |\nabla_y \tilde{I}_{\text{IR/VI}}|$. {Compute gradient magnitudes.}
  $\bar{G}_{ir/vi} \leftarrow \text{AvgPool}_{5\times5}(G_{ir/vi})$. {Suppress noise via average pooling.}
  $G^{\text{smooth}}_{ir/vi} \leftarrow \mathcal{G}_\sigma(G_{ir/vi})$. {Apply 3* 3 Gaussian smoothing.}
Step 3: Adaptive thresholding
  $\mathcal{P} \leftarrow \bar{G}_{ir} \cup \bar{G}_{vi}$. {Compute joint gradient pool.}
  $\theta \leftarrow Q_{\tau_q}(\mathcal{P})$. {Determine threshold: 80th percentile.}
  $\mathbf{M}_f \leftarrow \mathbb{1}[\bar{G}_{ir} < \theta] \wedge \mathbb{1}[\bar{G}_{vi} < \theta]$. {Build flatness mask.}
Step 4: Structure similarity scoring
  $\Delta G \leftarrow |G^{\text{smooth}}_{ir} - G^{\text{smooth}}_{vi}|_1$. {Compute gradient difference.}
  $S \leftarrow \exp(-\Delta G/\tau)$. {Compute similarity score cross modalities.}
  $\mathbf{M} \leftarrow \mathbb{1}\cdot[M_f] + \mathbb{1}\cdot[\neg M_f]\cdot S$. {Initialize mask.}
Step 5: Final refinement
  $\mathbf{M}_{\text{s}} \leftarrow \mathcal{G}_\sigma(\mathbf{M})$. {Apply Gaussian smoothing to suppress fringe artifacts.}

As Video Foundation Temporal Prior (VFTP), the key value of V-JEPA 2.0 lies in the fact that it learns the laws of temporal evolution in latent space, making it naturally suited to provide the temporal branch with a prior of 'how the world changes.' This complements the spatial branch, which requires a static-detail prior of 'what the scene looks like.' Notably, V-JEPA 2.0 is applied only to the visible frames, since it is pretrained on large-scale RGB video and its temporal prior is defined in the visible domain and feeding infrared frames would mismatch its pretrained input distribution. The infrared modality instead contributes complementary information through the spatial branch.

Since our visible input video has $T = 6$ frames while V-JEPA 2.0 expects 16 frames per clip, we repeat the visible video V_{VI} along the temporal dimension with repeats $[8, 8, 12, 12, 12, 12]$, yielding 64 frames that align with the temporal resolution of V-JEPA 2.0. Tab. 4 shows details of VideoMAE we used in the work.

A.2.2 VFSP: VIDEOMAE

VideoMAE Tong et al. (2022) is a self-supervised video pretraining framework based on the masked autoencoder (MAE) paradigm. It is the first to demonstrate that a pure ViT backbone can efficiently learn spatio-temporal representations from video data through masking-and-reconstruction alone. Its key innovations are the *Tube Masking* strategy and an extremely high masking ratio.

In our method, VideoMAE is adopted as the Video Foundation Spatial Prior (VFSP) for spatial fine-tuning. This is justified by its pretraining objective: VideoMAE performs *pixel-level reconstruction within each patch*, which forces the encoder to capture fine-grained spatial appearance details such as textures, edges, and local structures. As a result, its pretrained representation is naturally aligned with the spatial branch, which requires a static-detail prior of 'what the scene looks like.' This complements the temporal branch, which is modulated by V-JEPA 2.0 and focuses on 'how the world changes.' In contrast, V-JEPA 2.0 predicts latent representations of masked tokens, which emphasizes temporal dynamics rather than spatial appearance. The two priors therefore serve complementary roles in our decomposed dual-branch design.

The Tube Masking strategy of VideoMAE masks the same spatial location consistently across all frames, which prevents information leakage caused by temporal redundancy and forces the model to learn genuine spatio-temporal structure rather than simply copying adjacent frames. Its extremely high masking ratio (around 90%) performs best on video, in contrast to the 75% used in image MAE.

Table 4: Main parameters of V-JEPA 2.0 used as the Video Foundation Temporal Prior (VFTP).

Parameter	Value	Description
Hidden size	24	Feature width.
Hidden layers	24	Transformer blocks.
Attention heads	16	Per layer.
MLP ratio	4	FFN expansion ratio.
Input size	256×256	Spatial resolution.
Frames per clip	64	Frames per video clip.
Tubelet size	2	Temporal span of each spatio-temporal token.
Patch size	16	Spatial patch size.

Table 5: Main parameters of VideoMAE used as the Video Foundation Spatial Prior (VFSP).

Parameter	Value	Description
Hidden size	768	Encoder feature width.
Encoder layers	12	Transformer blocks.
Attention heads	12	Per encoder layer.
Intermediate size	3072	FFN expansion dimension.
Input size	224×224	Spatial resolution.
Frames	16	Frames per video clip.
Tubelet size	2	Temporal span of each spatio-temporal token.
Patch size	16	Spatial patch size.
Decoder hidden size	384	Lightweight decoder.
Decoder layers	4	Decoder Transformer blocks.
Decoder attention heads	6	Per decoder layer.
Norm pixel loss	true	MSE on normalized masked tokens.

In our dual-branch design, VideoMAE is well suited as the VFSP for three reasons. First, its pixel-level reconstruction directly optimizes appearance details, yielding fine-grained spatial representations that satisfy the spatial branch's need for a static-detail prior. Second, it is complementary to V-JEPA 2.0 because V-JEPA 2.0 predicts dynamics in latent space, while VideoMAE reconstructs appearance in pixel space, providing priors at different levels that match our decomposed structure. Third, it is a readily available pretrained model that can be directly used as a frozen spatial prior, avoiding the need for additional large-scale video pretraining. Tab.5 shows details of VideoMAE we used in the work.

Tab. 6 contrasts the two pretrained priors along prediction target, learning focus, masking strategy, and their role in our method. VideoMAE (VFSP) is trained with pixel-level reconstruction under tube masking, which drives it to capture spatial appearance details, whereas V-JEPA 2.0 (VFTP) predicts latent representations through joint embedding prediction and thus models temporal dynamics. This difference in pretraining objectives explains why the two priors specialize along different dimensions. Pixel-level reconstruction favors spatial fidelity, while prediction in latent space favors temporal reasoning. We therefore assign VideoMAE (VFSP) to the spatial branch for local detail refinement and V-JEPA 2.0 (VFTP) to the temporal branch for global motion modulation, so that each branch operates on the prior aligned with its objective.

Table 6: Comparison between VideoMAE (VFSP) and V-JEPA 2.0 (VFTP).

	VideoMAE (VFSP)	**V-JEPA 2.0 (VFTP)**
Prediction target	Pixel-level reconstruction	Latent representation prediction
Learning focus	Spatial appearance details	Temporal dynamics
Masking strategy	Tube Masking	Joint embedding prediction
Role in our method	Spatial branch modulation	Temporal branch modulation

A.3 DPP-BASED MECHANISM

Preliminaries. A Determinantal Point Process (DPP) defines a probability distribution over subsets $s \subseteq \mathcal{Y}$ through a positive semi-definite kernel matrix L:

$$\mathcal{P}(s) = \frac{\det(L_s)}{\det(L + \mathbb{I})},$$

where $L_s = [L_{ij}]_{i,j \in s}$ is the principal submatrix indexed by s. Given a quality score q_i and a similarity matrix K_{ij}, the DPP kernel is constructed as

$$L = \mathrm{diag}(q)\, K\, \mathrm{diag}(q); \qquad L_{ij} = q_i\, K_{ij}\, q_j.$$

The similarity matrix K is commonly instantiated as the cosine similarity between feature vectors.

$$K_{ij} = \frac{\langle f_i, f_j \rangle}{\|f_i\|\, \|f_j\|},$$

which is positive semi-definite and thus yields a valid DPP kernel. More generally, L only needs to be positive semi-definite; it is not restricted to a Gram matrix. The determinant $\det(L_s)$ serves as a joint measure of *quality* (via q_i) and *diversity* (via the off-diagonal structure of L). Maximizing $\det L_s$ favors a subset s that is simultaneously high-quality and mutually diverse, whereas minimizing it favors a subset that is mutually similar. This dual property underpins the two DPP modules in our method, which adopt two different kernel constructions.

Kernel Construction for Spatio-Temporal Consistency. In the unified spatio-temporal optimization, we enforce temporal consistency by selecting an arbitrary target step i and conducting optical flow warp from any previous step j ($j < i$) to i . We require the similarity between any pair of intermediate steps $j \to i$ and $k \to i$ to be as **high** as possible, i.e., we minimize their dissimilarity. To this end, we build the kernel matrix $\mathbf{T}$ from the temporal features using an ℓ_1 Laplacian kernel:

$$\mathbf{T}^i = \begin{pmatrix} t^i_{1,1} & t^i_{1,2} & \cdots & t^i_{1,i} \\ \vdots & \vdots & \ddots & \vdots \\ t^i_{j,1} & t^i_{j,2} & \cdots & t^i_{j,i} \\ \vdots & \vdots & \ddots & \vdots \\ t^i_{k,1} & t^i_{k,2} & \cdots & t^i_{k,i} \\ \vdots & \vdots & \ddots & \vdots \\ t^i_{i,1} & t^i_{i,2} & \cdots & t^i_{i,i} \end{pmatrix}_{i*i}, \quad t^i_{j,k} = \exp\Big(- \frac{|f^{t_j \to t_i}_{fu} - f^{t_k \to t_i}_{fu}|_1}{\sigma} \Big).$$

where $f^{t_j \to t_i}_{fu}$ and $f^{t_k \to t_i}_{fu}$ denote the warped temporal features from steps j and k to i. Since the Laplacian kernel is positive semi-definite, $\mathbf{T}$ is a valid DPP kernel. Moreover, the standard deviation of reconstructed frames before time step i from spatial branch feature f^s_{fu} is utilized to formulate the quality score diagonal matrix S^i. We then form the spatio-temporal DPP kernel L^{st}_{DPP} as:

$$(L^{st}_{DPP})^i = S^i\, \mathbf{T}^i\, (S^i)', \quad i = (1, 2, ..., T).$$

and **minimize** its determinant to ensure temporal coherence and enhance spatial details, achieving a trade-off between them so that the selected steps collapse toward a temporally consistent configuration.

$$\mathcal{L}^{st}_{DPP} = \sum_{i=1}^{T} \log\det\big(L^{st}_{DPP}\big)^i.$$

Kernel Construction for Cross-Modal Structural Alignment. For cross-modal structural alignment, we partition features into patches and consider the modality-specific features of infrared and visible light. Provided that the modality-specific features, combined with the shared features, are sufficient to reconstruct the input of each modality, we aim to increase the diversity between the two modalities' specific features so that each modality preserves its complementary information instead of collapsing into a shared representation. Accordingly, for the p-th non-overlapping patch,

the patch-level modality-specific representations $\tilde{f}_{ir}^{(p)}$ and $\tilde{f}_{vi}^{(p)}$ are used to construct a 2×2 cosine similarity matrix $\mathbf{K}^p$:

$$\mathbf{K}^p = \begin{pmatrix} \langle \tilde{f}_{ir}^{(p)}, \tilde{f}_{ir}^{(p)} \rangle & \langle \tilde{f}_{ir}^{(p)}, \tilde{f}_{vi}^{(p)} \rangle \\ \langle \tilde{f}_{vi}^{(p)}, \tilde{f}_{ir}^{(p)} \rangle & \langle \tilde{f}_{vi}^{(p)}, \tilde{f}_{vi}^{(p)} \rangle \end{pmatrix} = \begin{pmatrix} 1 & cos\theta_p \\ cos\theta_p & 1 \end{pmatrix},$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity. Since $\mathbf{K}^p$ is positive semi-definite, it is a valid DPP kernel. Its diagonal entries are fixed to 1 by construction, while the off-diagonal entry $cos\theta_p$ measures the redundancy between the two modalities. **Maximizing** the determinant of $\mathbf{K}^p$ therefore encourages the off-diagonal similarity to decrease, i.e., enlarges the diversity between infrared and visible specific features and prevents modality-specific information from being overwritten during alignment, thereby avoiding cross-modal representation collapse.

Since the reliability of the diversity constraint varies across patches, a confidence score q^p is further introduced to measure the gradient consistency between the fused image and the multi-modal reference in the p-th patch. A high q^p indicates a reliable patch that should contribute more to the diversity constraint, while a low q^p down-weights patches where the reconstruction is unreliable. To account for patch-wise reliability, a confidence weight q^p is further introduced to form the weighted kernel $\mathbf{Q}^p\mathbf{K}^p(\mathbf{Q}^p)'$, and the cross-modal DPP loss over all N patches is formulated as:

$$\mathcal{L}_{DPP}^{cm} = -\frac{1}{N}\sum_{p=1}^{N} \log\det\big(\mathbf{Q}^p\mathbf{K}^p(\mathbf{Q}^p)' + \epsilon \cdot \mathbf{I}\big),$$

where **maximizing** the determinant of $\mathbf{Q}^p\mathbf{K}^p(\mathbf{Q}^p)'$ is equivalent to minimizing $\mathcal{L}_{DPP}^{cm}$.

Summary. The two DPP modules are complementary: the spatio-temporal DPP loss $\mathcal{L}_{DPP}^{st}$ uses a ℓ_1 Laplacian kernel $\mathbf{T}$ for temporal coherence measurement and minimizes its determinant to enforce spatio-temporal consistency, whereas the cross-modal alignment DPP uses a cosine-similarity kernel and maximizes its determinant to preserve modality-specific diversity. Both kernels are positive semi-definite, and they promote consistency and complementarity in D^2Fusion.

A.4 Supplementary Experiments

A.4.1 Qualitative Comparisons

We show more visualization comparisons among baselines and D^2Fusion under distinct scenarios, including upstairs, crossroad, grassland, and snack stall in the supplementary materials. Moreover, in the supplementary materials, we also reflect performance under scenarios with different lighting intensities, including day and night. Existing IVIF methods such as LRRNet, GIFNet, and TIMFusion fail to significantly highlight infrared targets. Among IVVF methods, although Temcoco highlights infrared targets, it does not produce clear background details, while VideoFusion exhibits noticeable motion artifacts. In contrast, our method simultaneously retains clearer targets extracted from the infrared modality and preserves clearer visible-light background details without noticeable motion artifacts, demonstrating the superiority of our method.

A.4.2 Downstream Performance

Fig. 7 compares the object tracking performance of IVIF and IVVF baselines. For a fair comparison, all methods adopt YOLO v11 as the baseline tracker. In nighttime sports scenes, however, existing baselines exhibit three typical failure modes: incorrect target, missed tracking, and incorrect tracking. Incorrect target refers to the tracker locking onto an incorrect object; for instance, LRRNet and TIMFusion mistakenly identify the lights of a distant building as the tracked target, whereas D^2Fusion consistently maintains the correct target. Missed tracking denotes the failure to produce a track for a target that is actually present; for example, in frame 160, only D^2Fusion successfully tracks the player running on the right side of the screen, while all baselines lose track. Incorrect tracking refers to tracking a target with an incorrect identity, i.e., identity switches or cross-track contamination; for example, in frame 140, Temcoco erroneously tracks the person standing on the sidelines, while D^2Fusion correctly preserves the player's identity. These results demonstrate that D^2Fusion substantially outperforms existing baselines in challenging nighttime sports scenes. (See more scenario cases for object tracking in the supplementary materials.)

Table 7: Comparison of computation efficiency. '*' denotes IVIF methods which require extra deblurring cost under degraded scenarios.

Methods	LRRNet[2]	GIFNet*	TIMFusion*	DDFM*	Temcoco	VideoFusion	D^2Fusion
Time (s)	0.075	0.101	0.057	39.473	0.036	0.107	0.186
Flops (G)	14.170	212.954	138.229	5250.492	452.843	274.598	822.928

A.4.3 Computation Efficiency

As reported in Tab. 7, we evaluate the inference time and FLOPs of all compared methods on an A100 GPU with $T = 6$. D^2Fusion achieves competitive efficiency while additionally performing forward passes of VFTP V-JEPA 2.0 and VFSP VideoMAE, which are introduced to supply physical-world motion priors. To keep this overhead bounded, we run both models in FP16 and freeze all their weights without any fine-tuning, which reduces inference cost and simultaneously eliminates training overhead and the risk of overfitting on small fusion datasets. Many competing methods marked with '*' are designed for image-level fusion (IVIF) and therefore require an additional deblurring stage in degraded scenarios. This preprocessing overhead is not reflected in their reported fusion-network FLOPs, whereas IVVF methods (Temcoco, VideoFusion, and D^2Fusion) cover the full pipeline. Among all methods, DDFM is the most expensive because it performs multi-step (100-step) denoising. More importantly, our ablation study shows that removing V-JEPA 2.0 and VideoMAE as the VFTP and VFSP causes Temcoco to lose background details and VideoFusion to exhibit motion artifacts, since neither can extract motion priors of the physical world from large-scale video models. This confirms that the additional computation is not redundant but essential. Our design targets a better accuracy–efficiency trade-off: the extra cost buys explicit motion and physical priors that image-level fusion methods cannot provide. We deliberately adopt large-scale video models because they offer generalizable motion priors that lightweight alternatives fail to capture, as evidenced by the degraded results of Temcoco and VideoFusion.



Figure 7: Visualization comparison on object tracking downstream task.