SPDVF: UNIFIED INFRARED AND VISIBLE VIDEO FUSION ON LOG-EUCLIDEAN SYMMETRIC POSITIVE DEFINITE MANIFOLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing infrared and visible video fusion methods treat spatial fusion and temporal modeling as separate stages in independent feature spaces. Temporal evolution is thus decoupled from fusion decisions, leaving flickering artifacts in the fused sequence. We observe that the second-order statistics of local features encode cross-modal complementarity and temporal dynamics at once, so a single descriptor can serve both roles. Such statistics take the form of symmetric positive definite (SPD) matrices, which reside on a curved manifold where linear operations are no longer valid. We propose SPDVF, a video fusion framework that describes local regions by SPD covariance descriptors and, under the Log-Euclidean metric that flattens this manifold, performs cross-modal alignment and temporal propagation in one shared tangent space without explicit motion estimation. Three components realize this design. Log-Euclidean Cross-Modal Attention aligns dual-modal second-order statistics through shifted-window cross-attention. The Basis-Kernel Bridge then converts the fused tokens into kernel mixing weights that refine Euclidean detail features. Log-Euclidean Mamba finally propagates temporal state over the same fused SPD tokens at a cost linear in clip length. Experiments on M3SVD, HDO, and VTMOT show that SPDVF improves spatial fidelity while maintaining temporal consistency.

1 INTRODUCTION

In visual perception, a single imaging modality often fails to capture all relevant scene information. Visible sensors provide texture and color but degrade under adverse lighting, and infrared sensors highlight thermal targets yet lack fine-grained detail Bao et al. (2023). Combining these complementary sources through multi-modal fusion has therefore become essential for robust scene understanding Karim et al. (2023); Zhang et al. (2021); Zhang & Demiris (2023). Infrared and visible video fusion extends this idea to the temporal domain, seeking to produce fused video sequences that simultaneously exhibit enhanced target saliency, rich texture fidelity, and temporal continuity across frames. Fused video of this quality directly benefits downstream tasks such as intelligent surveillance, autonomous navigation, and target tracking in complex environments Muller & Narayanan (2009); Jain et al. (2023); Zhang et al. (2023; 2018).

This temporal extension, however, is far from a trivial per-frame generalization Zhao et al. (2026); Tang et al. (2025). Adjacent fused frames must transition smoothly in brightness, contrast, and tone, otherwise the output exhibits visible flickering, and because real-world video contains both camera ego-motion and independent object motion, the fusion process must perceive and adapt to dynamic scene changes rather than treat each frame in isolation. Temporal modeling must also scale gracefully with clip length, since designs whose cost grows super-linearly in the number of frames become impractical as sequences lengthen. These factors indicate that moving from image to video fusion is not a matter of applying a frame-level method repeatedly, but calls for rethinking how spatial fusion and temporal modeling should be coupled.

Recent works address temporal modeling through two strategies. The first warps neighboring frame features toward the current frame by optical flow before cross-modal interaction Zhao et al. (2026), the fusion decision itself remains frame-centric, and flow errors propagate directly as ghosting or

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.

blurring. The second encodes clips as spatio-temporal volumes via 3D convolutions or temporal attention Gong et al. (2025), capturing inter-frame dependencies more holistically but tending to over-smooth high-frequency details and scaling poorly with sequence length. Both share a deeper limitation: the temporal and fusion modules operate in separate feature spaces with independent parameterizations, so temporal consistency is pursued as a post-hoc regularization rather than emerging from the fusion representation itself. This leaves a clear gap: no existing method grounds cross-modal fusion and temporal propagation in a single shared representation, where both are constrained by the same structure.

Second-order statistics offer such a representation. Within a local region, the covariance of feature channels records how feature dimensions co-vary rather than how strongly each responds, which is what separates a thermally uniform hot target from a texture-rich background. The same descriptor drifts as the region moves or changes appearance, so its statistical signature also serves as its temporal state. After positive-definite regularization, these covariance descriptors are symmetric positive definite (SPD) and form a curved Riemannian manifold, on which the linear operations that fusion and state-space propagation both rely are no longer valid. The Log-Euclidean metric Arsigny et al. (2006) resolves this: the matrix logarithm maps each SPD descriptor diffeomorphically onto the tangent space at the identity, a flat space where dot-product attention and a linear state recurrence are again well defined. Mapping all local descriptors into this common tangent-space representation lets both processes operate on the same linearized geometry.

We instantiate this idea as SPDVF. Each local region becomes an SPD covariance token, and each token is mapped by a local LogEig operation into a shared Log-Euclidean tangent-space representation, where both fusion and temporal propagation act without repeated projection between the manifold and its tangent space. For intra-frame fusion, a Log-Euclidean Cross-Modal Attention (LECA) module aligns dual-modal second-order statistics via shifted-window cross-attention in the tangent space, and a Basis-Kernel Bridge (BKB) turns the fused statistical tokens into geometry-aware coefficients that refine the Euclidean detail features. For inter-frame modeling, a Log-Euclidean Mamba (LEM) module scans the fused SPD-token sequences bidirectionally at the deepest scale, at a cost linear in clip length, using local scan-transition differences as a lightweight change cue. Because LEM operates on the representations LECA produces, temporal evolution is informed by the statistical patterns that drive fusion decisions. Experiments on M3SVD, HDO, and VTMOT show consistent gains in spatial fidelity alongside competitive temporal stability. Our contributions are as follows:

- We propose SPDVF, a video fusion framework built on Log-Euclidean SPD covariance descriptors, where local LogEig mappings place the descriptors in a shared tangent-space representation so that cross-modal fusion and temporal modeling use the same geometry, tying how a region evolves to how it is fused.
- We design a Log-Euclidean Cross-Modal Attention module that aligns dual-modal second-order statistics by shifted-window cross-attention in the tangent space, paired with a Basis-Kernel Bridge that converts the fused descriptors into geometry-aware kernel-mixture coefficients to refine Euclidean detail features.
- We introduce a Log-Euclidean Mamba module that scans the fused descriptor sequences bidirectionally at a cost linear in both token count and clip length, and augment the serialized sequence with local scan differences to expose inter-frame transitions without optical flow.

2 Related Work

2.1 Multi-Modality Image Fusion

Learning-based infrared and visible image fusion has progressed along several architectural lines. Auto-encoder designs Li & Wu (2018); Zhao et al. (2020) learn a shared latent space and combine codes inside it, cross-attention architectures Li & Wu (2024); Wang et al. (2025a); Ma et al. (2022) compute the correspondence between modalities explicitly, and a related line trains a single network to cover several fusion scenarios at once Xu et al. (2020); Cheng et al. (2025). GAN- and diffusion-based methods Zhao et al. (2023b); Yi et al. (2024); Sun et al. (2026) recast fusion as distribution matching, which sharpens perceptual detail but offers less direct control over what is preserved.

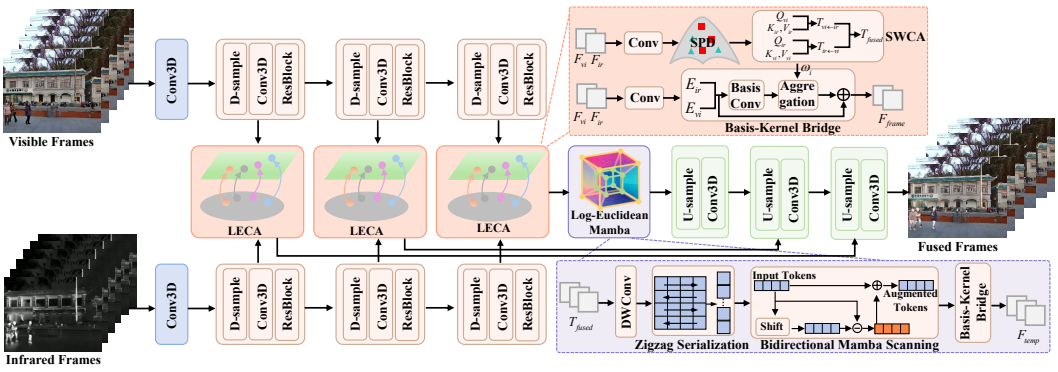


Figure 1: Overall pipeline of the proposed SPDVF. LECA denotes Log-Euclidean Cross-Modal Attention and SWCA denotes shifted-window cross-attention.

Later work relaxes the assumptions these methods make about their inputs, handling misalignment by joint registration and fusion Wang et al. (2022) or degraded sources by suppressing noise and blur along the way Wang et al. (2025b), while task-driven designs let a downstream detector or segmenter decide what matters Zhao et al. (2023a); Liu et al. (2024a). What none of them revisits is the arithmetic of the merge itself, which stays first-order throughout: pixel- and channel-level correspondence is resolved, but how feature dimensions co-vary within a region is not. Covariance descriptors capture that, at the price of living on a curved manifold, and SMLNet Kang et al. (2025) takes this route for static image pairs. All of these methods operate on single frames, however, and carry no notion of how a fusion decision should evolve over time.

2.2 Multi-Modality Video Fusion

Applying an image fusion method frame by frame leaves each frame's decision independent of the previous one, and the resulting brightness jitter appears as flicker. Two families of fixes have emerged, differing in where temporal information enters the pipeline. The first aligns before it fuses. UniVF Zhao et al. (2026) warps neighboring frames onto the current one by optical flow and adds a temporal consistency loss on top. Alignment quality then becomes the bottleneck: flow is unreliable under occlusion and fast motion, and because each frame pair must be solved in turn, the cost is latency-bound rather than arithmetic-bound. Temporal information has also done its work before fusion begins, leaving the fusion decision itself frame-centric. The second fuses the clip as a volume, through 3D convolution or temporal attention. VideoFusion Tang et al. (2025) aggregates forward and backward context with a bi-temporal co-attention, and SDMFusion Liu et al. (2026) pairs a one-step streaming diffusion sampler with a memory-augmented latent pipeline to keep this affordable at video rates. Aggregating densely over space and time captures dependencies the warping methods miss, but it also averages high-frequency detail across frames, and the cost grows faster than the clip does.

3 Method

3.1 Overview

Given infrared and visible sequences $\mathbf{V}_{ir}, \mathbf{V}_{vi} \in \mathbb{R}^{B\times T\times C\times H\times W}$, SPDVF produces a fused video $\mathbf{Y}_f \in \mathbb{R}^{B\times T\times 3\times H\times W}$ that integrates cross-modal complementary information while remaining temporally coherent. Fig. 1 shows the pipeline: a three-scale U-shaped backbone whose modality-specific 3D convolutions yield features $\mathbf{F}^n_{ir}, \mathbf{F}^n_{vi} \in \mathbb{R}^{T\times C_n\times H_n\times W_n}$ at scales $n \in \{1, 2, 3\}$. A Log-Euclidean Cross-Modal Attention (LECA) module fuses the two modalities at every scale, and a Log-Euclidean Mamba (LEM) module propagates the fused tokens across frames at the deepest scale, where temporal drift appears as brightness and tone shifts of coherent regions that the coarsest descriptors already summarize, shallow features reach the decoder through skip connections. Two branches run in parallel throughout: the SPD branch decides where and how to fuse from second-

order statistics, the Euclidean branch retains the fine detail needed for reconstruction, and a basis-kernel bridge couples them.

3.2 PRELIMINARIES

We state the geometry sketched above precisely. The set of $d \times d$ SPD matrices Sym^d_{++} forms a Riemannian manifold, and standard linear operations are not valid on this curved space. Under the Log-Euclidean metric Arsigny et al. (2006), the matrix logarithm $\log : \mathrm{Sym}^d_{++} \to \mathrm{Sym}^d$ is a *global diffeomorphism* that maps the manifold onto the tangent space at the identity, which is a flat vector space, thereby converting the multiplicative structure of SPD matrices into an additive one. Geodesic distances then take the closed form $\delta_{LE}(\mathbf{C}_1, \mathbf{C}_2) = \| \log \mathbf{C}_1 - \log \mathbf{C}_2 \|_F$, and convex combinations $\sum_j a_j \log \mathbf{C}_j$ map back through $\exp(\cdot)$ to weighted means on the manifold. Applying this mapping to each local descriptor places all tokens in one fixed tangent space, where the additive structure makes standard Euclidean operations well defined for both cross-modal and temporal processing.

3.3 SPD REPRESENTATION CONSTRUCTION

For each frame t at scale n, the 3D encoder produces $\mathbf{F}^{t,n} \in \mathbb{R}^{C_n \times H_n \times W_n}$, whose channels are reduced from C_n to d by a 1×1 convolution, yielding $\tilde{\mathbf{F}}^{t,n} \in \mathbb{R}^{d \times H_n \times W_n}$. We partition the spatial domain into $N_w = \lfloor H_n/p \rfloor \times \lfloor W_n/p \rfloor$ non-overlapping $p \times p$ windows. For window i, the $N = p^2$ positions are stacked into a local feature matrix $\mathbf{Z}_i \in \mathbb{R}^{N \times d}$. Its sample covariance is:

$$\hat{\mathbf{C}}_i = \frac{1}{N-1} \sum_{k=1}^{N} (\mathbf{z}_{k,i} - \bar{\mathbf{z}}_i)(\mathbf{z}_{k,i} - \bar{\mathbf{z}}_i)^\top, \tag{1}$$

where $\mathbf{z}_{k,i}$ is the feature at position k and $\bar{\mathbf{z}}_i$ their mean. Because N is small, $\hat{\mathbf{C}}_i$ tends to be ill-conditioned, so we regularize it by shrinkage toward a scaled identity Ledoit & Wolf (2004):

$$\mathbf{C}_i = (1-\rho)\,\hat{\mathbf{C}}_i + \rho\,\frac{\mathrm{tr}(\hat{\mathbf{C}}_i)}{d}\,\mathbf{I}_d + \epsilon \mathbf{I}_d, \tag{2}$$

where $\rho \in (0,1)$ is the shrinkage coefficient, $\epsilon > 0$ is a fixed numerical floor, $\mathrm{tr}(\cdot)$ denotes the matrix trace, and $\mathbf{I}_d$ is the $d \times d$ identity matrix. The shrinkage term improves conditioning, while the positive floor ensures strict positive definiteness even when the sample covariance is rank deficient or zero; hence $\mathbf{C}_i \in \mathrm{Sym}^d_{++}$. We keep ρ and ϵ fixed rather than solving for a risk-optimal coefficient, since the $N = p^2$ positions in a window are spatially correlated and their distribution shifts as the encoder trains. This keeps the eigenvalue floor uniform across windows and training steps, improving the conditioning of the backward pass through the eigendecomposition. Each SPD matrix is then mapped into the tangent space via eigendecomposition $\mathbf{C}_i = \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^\top$ followed by a LogEig operation Huang & Van Gool (2017):

$$\mathbf{S}_i = \mathbf{U}_i \log(\mathbf{\Lambda}_i) \mathbf{U}_i^\top, \tag{3}$$

where $\mathbf{U}_i$ and $\mathbf{\Lambda}_i$ are the eigenvectors and eigenvalues of $\mathbf{C}_i$. The symmetric $\mathbf{S}_i$ is vectorized into a tangent-space token $\mathbf{t}_i \in \mathbb{R}^D$, $D = d(d+1)/2$, by stacking its diagonal with its upper-triangular entries and scaling the off-diagonal entries by $\sqrt{2}$, giving $\mathbf{t}_{vi,i}$ and $\mathbf{t}_{ir,i}$ per modality. Since each off-diagonal of a symmetric matrix appears twice, this weight makes the vectorization an isometry:

$$\langle \mathbf{t}_i, \mathbf{t}_j \rangle = \sum_a S^i_{aa} S^j_{aa} + 2 \sum_{a<b} S^i_{ab} S^j_{ab} = \langle \mathbf{S}_i, \mathbf{S}_j \rangle_F, \tag{4}$$

so that $\|\mathbf{t}_i - \mathbf{t}_j\|_2 = \delta_{LE}(\mathbf{C}_i, \mathbf{C}_j)$. Euclidean inner products and distances between tokens therefore coincide exactly with their Log-Euclidean counterparts on the manifold, and the tokens remain elements of a fixed linear space in which the state-space recurrence is well defined.

3.4 LOG-EUCLIDEAN CROSS-MODAL ATTENTION

LECA aligns dual-modal second-order statistics in the tangent space and derives geometry-aware coefficients that refine Euclidean detail features. The Euclidean branch retains first-order details for

high-fidelity reconstruction: $\mathbf{E}_{vi} = \mathcal{F}_{conv}(\mathbf{F}_{vi})$ and $\mathbf{E}_{ir} = \mathcal{F}_{conv}(\mathbf{F}_{ir})$, where $\mathcal{F}_{conv}(\cdot)$ denotes a convolutional feature extractor and $\mathbf{F}_{vi}, \mathbf{F}_{ir}$ are the encoder features at the current scale. The SPD branch produces the tangent-space tokens $\mathbf{t}_{vi,i}, \mathbf{t}_{ir,i}$ by vectorizing $\mathbf{S}_i$ of Eq. equation 3, with the descriptor dimension d fixed by the 1×1 reduction.

Local Window Cross-Attention. We restrict cross-modal attention to local $k \times k$ windows of k^2 tokens. Within each window, the visible-to-infrared direction reads

$$\mathbf{T}_{vi \leftarrow ir} = softmax\left(\frac{\mathbf{Q}_{vi}\mathbf{K}_{ir}^{\top}}{\sqrt{D}}\right)\mathbf{V}_{ir}, \tag{5}$$

where $\mathbf{Q}_{vi} = \mathbf{T}_{vi}\mathbf{W}_Q$, $\mathbf{K}_{ir} = \mathbf{T}_{ir}\mathbf{W}_K$, $\mathbf{V}_{ir} = \mathbf{T}_{ir}\mathbf{W}_V \in \mathbb{R}^{k^2 \times D}$, $\mathbf{T}_{vi}, \mathbf{T}_{ir} \in \mathbb{R}^{k^2 \times D}$ are the stacked token matrices within each window, and $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{D \times D}$ are learnable projections. By Eq. equation 4, inner products and distances between the tokens $\mathbf{T}_{vi}, \mathbf{T}_{ir}$ coincide with the Frobenius inner product and the geodesic distance δ_{LE} of the corresponding descriptors. The projections therefore act on a representation whose metric already reflects the manifold geometry, and the affinity $\mathbf{Q}_{vi}\mathbf{K}_{ir}^{\top}$ is a learned bilinear form on that representation rather than on an arbitrary rescaling of the second-order statistics. The counterpart $\mathbf{T}_{ir \leftarrow vi}$ is computed analogously, and the fused tokens are $\mathbf{T}_{fused} = (\mathbf{T}_{vi \leftarrow ir} + \mathbf{T}_{ir \leftarrow vi})/2$.

Shifted-Window Interaction and Complexity. Following the shifted-window scheme of hierarchical vision transformers Liu et al. (2021), odd-numbered layers shift the partition by $\lfloor k/2 \rfloor$ positions so that boundary tokens interact in alternating layers, allowing information to flow across window boundaries. Restricting attention to local windows yields $O(N_w\, k^2)$ complexity, linear in the number of SPD tokens, compared with $O(N_w^2)$ for global attention, which would additionally require pairwise geodesic-distance computations on the manifold. The representation is not free, however: forming the N_w descriptors and mapping them through the matrix logarithm costs a batched $O(N_w d^3)$ eigendecomposition, an overhead first-order pipelines avoid. Since d is small and fixed by the 1×1 reduction, this term stays bounded as the spatial resolution grows.

Basis-Kernel Bridge (BKB). The tangent-space decisions must return to the Euclidean reconstruction space. We keep a bank of K shared depthwise basis kernels $\{\mathbf{\Phi}_j\}_{j=1}^{K}$ and let each fused token predict their mixing weights, which avoids both extremes: predicting a full kernel per window would be costly, while a single channel-attention vector would collapse the statistics into a first-order summary. For window i, the fused token $\mathbf{t}_{fused,i}$ predicts coefficients through a learnable projection $\mathbf{W}_c$ and a softmax:

$$\boldsymbol{\omega}_i = softmax(\mathbf{W}_c\, \mathbf{t}_{fused,i}) \in \mathbb{R}^K. \tag{6}$$

Meanwhile, the Euclidean features are combined, $\mathbf{E} = \mathbf{E}_{vi} + \mathbf{E}_{ir}$, and convolved with every basis kernel to form response maps $\mathbf{R}_j = \mathbf{\Phi}_j \circledast \mathbf{E}$, where $\circledast$ denotes depthwise convolution. Expanding the per-window coefficients $\boldsymbol{\omega}_i$ into spatial weight maps $\mathbf{\Omega}_j$, the output combines the shared responses with a residual fusion path:

$$\mathbf{F}_{frame} = \mathcal{P}\left(\sum_{j=1}^{K} \mathbf{\Omega}_j \odot \mathbf{R}_j\right) + \mathbf{W}_f\,[\mathbf{E}_{vi}; \mathbf{E}_{ir}], \tag{7}$$

where $\mathcal{P}$ and $\mathbf{W}_f$ are 1×1 convolutions, $\odot$ denotes element-wise multiplication, and $[\cdot\,;\cdot]$ denotes channel-wise concatenation. The geometry-aware coefficients let each region weight how second-order statistics shape its detail features, while the residual path provides a direct fusion channel. The same bridge structure (with an independent head) is reused in LEM, giving both stages a consistent manifold-to-Euclidean interface.

3.5 Log-Euclidean Mamba

To capture the temporal evolution of fused statistical patterns at a cost linear in clip length, LEM performs bidirectional selective state-space scanning Gu & Dao (2023) on the fused SPD-token sequences.

Shared Token Sequence. LECA already outputs tangent-space tokens $\mathbf{t}_{fused,i}^{t} \in \mathbb{R}^D$, which we take directly as LEM input, $\mathbf{x}_i^t := \mathbf{t}_{fused,i}^{t}$, with no re-encoding and no further $\log$ / $\exp$. Since the single identity-based logarithm has already linearized the manifold, the linear state recurrence and

affine updates of a selective SSM are valid tangent-space operations, and inter-frame state evolution is thus constrained by the same second-order statistics that drove the fusion decision.

Zigzag Serialization. A per-window token tracking a moving target may cross window boundaries over time, so we serialize the 2D grid while preserving spatial adjacency. Let $\mathbf{X}^t \in \mathbb{R}^{N_w \times D}$ denote the grid of SPD tokens of frame t. A depthwise convolution first blends each token with its spatial neighbors,

$$\tilde{\mathbf{X}}^t = \mathbf{X}^t + \mathcal{F}_{dw}(\mathbf{X}^t), \tag{8}$$

where $\mathcal{F}_{dw}(\cdot)$ denotes a depthwise 3×3 convolution. The tokens of each $\tilde{\mathbf{X}}^t$ are then flattened in zigzag order (odd rows left-to-right, even rows right-to-left) and concatenated across frames in temporal order, forming a single sequence of length $T \cdot N_w$. This ordering preserves local spatial continuity within each frame while allowing the recurrent state to propagate across successive frame blocks.

Shift-Based Difference Augmentation. Let $\tilde{\mathbf{x}}_k \in \mathbb{R}^D$ denote the k-th token vector of the flattened sequence. We augment the input with a local difference along the serialized scan before scanning:

$$\tilde{\mathbf{x}}_k^{aug} = \tilde{\mathbf{x}}_k + \eta\,(\tilde{\mathbf{x}}_k - \tilde{\mathbf{x}}_{k-1})\,, \tag{9}$$

where η is a learnable scalar initialized to a small positive value and $\tilde{\mathbf{x}}_0 = \tilde{\mathbf{x}}_1$. The difference term covers both spatial changes within a frame and transitions between consecutive frame blocks, so it is a scan-transition cue rather than an aligned motion estimate. Feeding it through the input-dependent discretization of the selective SSM lets the cue shape the state update without touching the standard selective-scan implementation, and the relation between transition magnitude and update strength is learned end-to-end rather than prescribed.

Bidirectional Mamba Scanning. Following bidirectional visual state-space designs Zhu et al. (2024); Liu et al. (2024b), we apply bidirectional selective scanning on the augmented sequence:

$$\mathbf{h}_k = \bar{\mathbf{A}}_k \mathbf{h}_{k-1} + \bar{\mathbf{B}}_k \tilde{\mathbf{x}}_k^{aug}, \quad \mathbf{y}_k = \mathbf{C}_k^{\text{ssm}} \mathbf{h}_k, \tag{10}$$

where $\bar{\mathbf{A}}_k$, $\bar{\mathbf{B}}_k$, $\mathbf{C}_k^{\text{ssm}}$ are input-dependent discretized parameters, $\mathbf{h}_k \in \mathbb{R}^D$ is the hidden state, and $\mathbf{y}_k$ is the output at position k. The forward and backward scans aggregate complementary context from the two serialization directions. The two directions are averaged, yielding symmetric spatio-temporal aggregation.

Basis-Kernel Temporal Projection. The scan output remains a sequence of Log-Euclidean tokens. These are transferred back to the Euclidean space through the same Basis-Kernel Bridge as LECA (with an independent head). The resulting temporal modulation $\mathbf{F}_{temp}$ is added to the intra-frame fused feature as a residual:

$$\mathbf{F}_{frame}^{out} = \mathbf{F}_{frame} + \mathbf{F}_{temp}, \tag{11}$$

where $\mathbf{F}_{temp}$ denotes the temporal modulation from the scan output. Temporal modeling thus augments the fused representation while leaving its spatial content intact.

3.6 Loss Function

The total loss jointly constrains spatial quality and temporal consistency:

$$\mathcal{L}_{total} = \mathcal{L}_{fuse} + \lambda_1 \mathcal{L}_{fid} + \lambda_2 \mathcal{L}_{tc}, \tag{12}$$

where λ_1, λ_2 are balancing coefficients. The fusion loss $\mathcal{L}_{fuse}$ combines structural, intensity, gradient, and color terms against both sources; the fidelity loss $\mathcal{L}_{fid}$ reconstructs the source frames from $\mathbf{F}_{frame}$ through two lightweight auxiliary decoders that are discarded at inference; and the temporal consistency loss $\mathcal{L}_{tc}$ aligns the inter-frame variation direction of the fused video with both sources, complementing the tangent-space modeling in LEM by suppressing flickering in the pixel domain without optical flow. Full definitions are given in Appendix A.

4 Experiments

4.1 Experiment Settings

To validate the effectiveness of the proposed method, we select seven representative fusion methods for comparison, including three image fusion methods (GIFNet, TIMFusion, and CtrlFuse) and four

Method	M3SVD				HDO				VTMOT			
	$MI\uparrow$	$VIF\uparrow$	$BiSwE\downarrow$	$MS2R\downarrow$	$MI\uparrow$	$VIF\uparrow$	$BiSwE\downarrow$	$MS2R\downarrow$	$MI\uparrow$	$VIF\uparrow$	$BiSwE\downarrow$	$MS2R\downarrow$
TIMFusion	2.9179	0.5641	2.3187	0.1288	2.6178	0.5710	4.8887	0.0559	2.4501	0.4908	7.8969	<u>0.6734</u>
GIFNet	2.8378	0.5655	2.8329	0.1221	2.4010	0.5490	7.1629	0.0527	2.2318	0.5796	13.4454	0.7029
CtrlFuse	4.4528	0.8535	2.9514	0.1207	2.4885	0.6586	5.2588	0.0430	2.3825	<u>0.6280</u>	9.4471	0.6981
TemCoCo	3.4293	0.6404	5.7624	0.6925	2.4057	0.5175	4.3021	0.0750	2.4607	0.5141	9.2848	0.9184
UniVF	<u>5.3733</u>	<u>0.9815</u>	2.7929	<u>0.1157</u>	<u>3.1129</u>	<u>0.6970</u>	3.8294	<u>0.0403</u>	2.1251	0.4433	9.1294	0.8967
VideoFusion	4.1464	0.7584	**1.4999**	0.1391	3.0805	0.5660	**2.2060**	0.0647	<u>3.0768</u>	0.5053	6.1611	1.3348
CMVF	4.1577	0.8255	2.3395	0.1309	2.7564	0.6772	3.3883	0.0415	2.3146	0.5704	<u>5.7992</u>	0.7050
Ours	**5.5801**	**0.9867**	<u>2.2712</u>	**0.1152**	**3.2357**	**0.7060**	<u>3.2323</u>	**0.0402**	**4.2148**	**0.7235**	**4.8656**	**0.6564**

Table 1: Quantitative comparison on M3SVD, HDO, and VTMOT datasets. **Bold** indicates the best, and underline indicates the second best indicates the second best. ↑ means higher is better and ↓ means lower is better.

video fusion methods (TemCoCo, UniVF, VideoFusion, and CMVF). Experiments are conducted on three publicly available video fusion benchmark datasets: M3SVD Tang et al. (2025), HDO Xie et al. (2024), and VTMOT Zhao et al. (2026). For each dataset, we randomly select 5 video sequences for testing, with the remaining sequences used for training. The proposed model is implemented in PyTorch and trained on two NVIDIA RTX A6000 GPUs using distributed data parallel. The input patch size is 128×128 with a batch size of 4. We employ the AdamW optimizer with an initial learning rate of 5×10^{-5}, a weight decay of 1×10^{-5}. The entire training process consists of 10 epochs. The hyper-parameters balancing the loss terms are empirically set as $\lambda_1 = 1$, $\lambda_2 = 10$, $\alpha_1 = 15$, $\alpha_2 = 1$, and $\alpha_3 = 10$.

4.2 Results of Multi-Modality Video Fusion

In this section, we compare the performance of the proposed method with 3 state-of-the-art image fusion methods, including TIMFusion Liu et al. (2024a), GIFNet Cheng et al. (2025), CtrlFuse Sun et al. (2026) and four video fusion methods, TemCoCo Gong et al. (2025), UniVF Zhao et al. (2026), VideoFusion Tang et al. (2025), CMVF Ding et al. (2026).

Quantitative Comparisons. We adopt two spatial metrics, MI and VIF, and two temporal metrics, BiSwE and MS2R Zhao et al. (2026), for evaluation. As shown in Table 1, SPDVF achieves the best results on 10 out of 12 metric-dataset combinations, ranking first in MI, VIF, and MS2R across all three datasets. On BiSwE, VideoFusion obtains the lowest scores on M3SVD and HDO owing to its dense temporal aggregation, while our method ranks a close second. On VTMOT, SPDVF leads in all four metrics. Compared with the strongest baseline UniVF, our method improves MI and VIF consistently while matching or surpassing its temporal scores without relying on optical flow estimation.

Qualitative Comparisons. Fig. 2 shows representative results from the three datasets. Image fusion methods suffer from reduced contrast or blurred edges in the absence of temporal modeling. VideoFusion preserves structural details but introduces ghosting near moving objects due to flow errors, while UniVF maintains smoothness at the cost of over-smoothed textures. Our method preserves both thermal saliency and fine-grained texture while maintaining consistent brightness across frames.

4.3 Evaluation of Temporal Consistency

To evaluate the temporal stability and continuity of fused videos, we visualize sequences from the M3SVD dataset by analyzing pixel-level intensity variations between consecutive frames. Specifically, we subtract each preceding frame from its successor to obtain frame-difference maps, with positive deviations shown in red and negative deviations in blue. In a temporally consistent video, static background regions should appear as pure white, meaning zero fluctuation, while moving objects leave clear, continuous color trails. As shown in Fig. 4, other methods such as VideoFusion produce scattered noise artifacts in the difference maps. In contrast, our method produces cleaner backgrounds and sharper motion trails, confirming its stronger temporal coherence.



Figure 2: Qualitative comparisons between our method and existing fusion approaches on the three datasets.

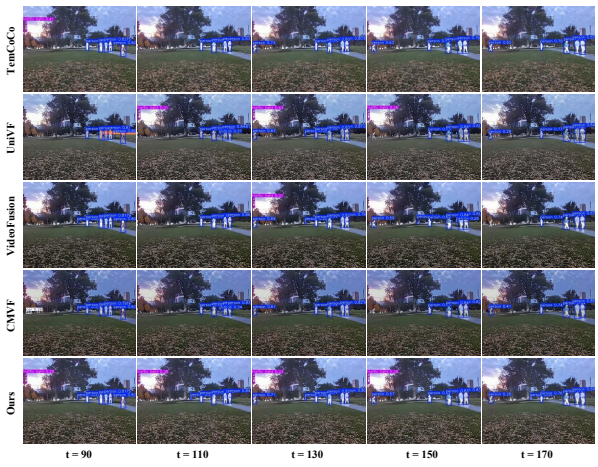


Figure 3: Qualitative comparison of object detection on M3SVD.



Figure 4: Temporal consistency visualization on M3SVD. Frame-difference maps are computed between consecutive frames.

4.4 Object Detection

Effective information aggregation benefits downstream machine vision tasks as well as visual perception. To examine what each fused video exposes to a detector, we apply a pre-trained YOLO v11 Khanam & Hussain (2024) to every fused result without task-specific fine-tuning and inspect which targets are boxed and how the boxes behave across adjacent frames. As shown in Fig. 3, thermally salient pedestrians that the low-contrast baselines leave unboxed are recovered in our results, and

Metric	*w/o* LECA	*w/o* LEM	*w/o* BKB	*w/* Decouple	Ours
$MI \uparrow$	5.2211	5.3127	5.1677	<u>5.5136</u>	**5.5801**
$VIF \uparrow$	0.9639	0.9764	0.9508	<u>0.9826</u>	**0.9867**
$BiSwE \downarrow$	<u>2.2479</u>	2.3555	**2.2403**	2.3419	2.2712
$MS2R \downarrow$	0.1194	<u>0.1175</u>	0.1205	0.1192	**0.1152**

Table 2: Ablation study on the M3SVD dataset.

Metric	Video-Fusion	UniVF	TemCoCo	Ours
Param (M)	6.75	9.16	19.21	23.62
FLOPs (G)	274.5	753.7	454.7	953.1
Time (ms)	141.0	302.9	83.6	237.9

Table 3: Efficiency comparison of video fusion methods on M3SVD.

the boxes persist over consecutive frames instead of flickering. Both effects follow from modeling cross-modal complementarity and temporal context on a shared statistical manifold, which retains thermal saliency without sacrificing the texture the detector relies on.

4.5 Ablation Study

We build four variants on M3SVD: three remove one module each (*w/o* LECA, *w/o* LEM, *w/o* BKB), while *w/* Decouple keeps all three but lets LEM read a separately re-encoded Euclidean token stream, testing the shared representation rather than any single module. As Table 2 shows, every removal lowers MI and VIF while raising MS2R, with BKB costing the most in detail fidelity. BiSwE splits by module type: removing LECA or BKB lowers it, since BiSwE compares a sequence against warped copies of itself with no reference to the sources and thus rewards outputs carrying less high-frequency content, whereas removing LEM, the only temporal module, raises it as expected. The decoupled variant replaces the shared fused descriptors with a separately re-encoded Euclidean token stream as input to LEM. The resulting degradation in both temporal metrics indicates that sharing the fused representation primarily improves temporal coherence.

4.6 Efficiency Analysis

As Table 3 shows, SPDVF has the largest parameter count and FLOPs of the four methods, and is slower than VideoFusion though faster than UniVF, so the gains of Table 1 are not obtained at parity of compute. The two costs differ in origin. Parameters are dominated by the backbone, a width-doubled VideoFusion that alone reaches 26.78M against our 23.62M, while the manifold components add 1.50M, or 6.3%. FLOPs and latency, by contrast, are attributable to the manifold branch: the batched covariance construction and eigendecomposition cost $O(N_w d^3)$, which first-order pipelines do not pay, and this is why temporal propagation is confined to the deepest scale, where N_w is smallest. Decoupling likewise costs both temporal metrics while single-frame.

5 Conclusion

This paper introduced SPDVF, a unified framework that performs infrared-visible video fusion and temporal modeling within a shared Log-Euclidean tangent space. SPDVF represents local regions using SPD covariance descriptors and aligns cross-modal second-order statistics through Log-Euclidean Cross-Modal Attention. The Basis-Kernel Bridge converts the fused descriptors into geometry-aware refinements of Euclidean features, while Log-Euclidean Mamba propagates temporal states over the same representation. Local differences along the serialized scans provide transition cues, removing the need for explicit optical-flow estimation. Extensive experiments on diverse video fusion datasets validate the superiority of our approach in both fusion quality and temporal stability.

AI use statement

In preparing this submission, we used generative AI tools, including ChatGPT, to assist with translation, literature discovery, drafting and editing selected passages, and LaTeX formatting. These tools also provided feedback on the presentation of the method, experimental results, figures, and tables. The authors reviewed all AI-assisted material and checked the final technical claims, citations, and numerical results against the relevant sources and experimental records. Generative AI tools were

not used to create the benchmark data or execute the reported experiments. The authors take full responsibility for the final content of this work, including AI-assisted text and artifacts.

Ethics statement

This work evaluates DGP on the publicly available MM-IMDb, UPMC Food101, Hateful Memes, CMU-MOSI, and CMU-MOSEI benchmarks. It does not introduce a new dataset, recruit participants, or collect new personal data. Some benchmark content, particularly Hateful Memes, may be offensive, while the sentiment datasets include recordings of individuals. The data are used for academic evaluation subject to their applicable terms. DGP may inherit biases and coverage limitations from these datasets and pretrained encoders, and missing modalities may further affect performance across different populations or conditions. Benchmark results should not be interpreted as a guarantee of fairness or reliability in deployment.

Reproducibility statement

Section 3 describes the sensitivity-based prompt routing, paired cross-modal prototype memory, and training objectives. Section 4 specifies the datasets, evaluation metrics, missing-modality settings, backbone, and main implementation choices, including the tri-modal variant used for sentiment analysis. Appendix A provides additional training details and visual-recognition hyperparameters. Appendix B reports computational cost and stability across three random seeds under a representative missing-both setting.

References

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine*, 56(2):411–421, 2006.

Fanglin Bao, Xueji Wang, Shree Hari Sureshbabu, Gautam Sreekumar, Liping Yang, Vaneet Aggarwal, Vishnu N Boddeti, and Zubin Jacob. Heat-assisted detection and ranging. *Nature*, 619 (7971):743–748, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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.

Meiqi Gong, Hao Zhang, Xunpeng Yi, Linfeng Tang, and Jiayi Ma. Temcoco: Temporally consistent multi-modal video fusion with visual-semantic collaboration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14326–14335, 2025.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Zhiwu Huang and Luc Van Gool. A riemannian network for spd matrix learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 2036–2042, 2017.

Deepak Kumar Jain, Xudong Zhao, Germán González-Almagro, Chenquan Gan, and Ketan Kotecha. Multimodal pedestrian detection using metaheuristics with deep convolutional neural network in crowded scenes. *Information Fusion*, 95:401–414, 2023.

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

Shahid Karim, Geng Tong, Jinyang Li, Akeel Qadir, Umar Farooq, and Yiting Yu. Current advances and future perspectives of image fusion: A comprehensive review. *Information Fusion*, 90:185–217, 2023.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE transactions on image processing*, 28(5):2614–2623, 2018.

Hui Li and Xiao-Jun Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103:102147, 2024.

Jinyuan Liu, Ludan Sun, Tengyu Ma, Chunyan Yang, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Streaming diffusion model for fast infrared and visible video fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14305–14314, 2026.

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

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103031–103063, 2024b.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

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

Amanda C Muller and Sundaram Narayanan. Cognitively-engineered multisensor image fusion for military applications. *Information Fusion*, 10(2):137–149, 2009.

Yiming Sun, Yuan Ruan, Qinghua Hu, and Pengfei Zhu. Ctrlfuse: Mask-prompt guided controllable infrared and visible image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9260–9268, 2026.

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 and restoration. *arXiv preprint arXiv:2503.23359*, 2025.

Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. *arXiv preprint arXiv:2205.11876*, 2022.

Dongliang Wang, Chuang Huang, Hao Pan, Yuan Sun, Jian Dai, Yanan Li, and Zhenwen Ren. Amlca: Additive multi-layer convolution-guided cross-attention network for visible and infrared image fusion. *Pattern Recognition*, 163:111468, 2025a.

Xue Wang, Zheng Guan, Wenhua Qian, Jinde Cao, Runzhuo Ma, and Cong Bi. A degradation-aware guided fusion network for infrared and visible image. *Information Fusion*, 118:102931, 2025b.

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*, 26:11031–11043, 2024.

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

Xunpeng Yi, Linfeng Tang, Hao Zhang, Han Xu, and Jiayi Ma. Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110:102450, 2024.

Hao Zhang, Han Xu, Xin Tian, Junjun Jiang, and Jiayi Ma. Image fusion meets deep learning: A survey and perspective. *Information Fusion*, 76:323–336, 2021.

Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12):14679–14694, 2023.

Xingchen Zhang and Yiannis Demiris. Visible and infrared image fusion using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):10535–10554, 2023.

Yue Zhang, Bin Song, Xiaojiang Du, and Mohsen Guizani. Vehicle tracking using surveillance with multimodal data fusion. *IEEE Transactions on Intelligent Transportation Systems*, 19(7): 2353–2361, 2018.

Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13955–13965, 2023a.

Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Pengfei Li, and Jiangshe Zhang. Didfuse: Deep image decomposition for infrared and visible image fusion. *arXiv preprint arXiv:2003.09210*, 2020.

Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, Kai Zhang, Deyu Meng, Radu Timofte, and Luc Van Gool. Ddfm: Denoising diffusion model for multi-modality image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8082–8093, 2023b.

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. *Advances in Neural Information Processing Systems*, 38:128504–128535, 2026.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *Proceedings of the International Conference on Machine Learning*, pp. 62429–62442, 2024.

A Loss Function Details

Fusion Loss. The fusion loss combines structural, intensity, gradient, and color terms:

$$\mathcal{L}_{fuse} = \mathcal{L}_{ssim} + \alpha_1 \mathcal{L}_{int} + \alpha_2 \mathcal{L}_{grad} + \alpha_3 \mathcal{L}_{color}, \tag{13}$$

where $\mathcal{L}_{ssim}$ measures structural similarity with both sources, and $\alpha_1, \alpha_2, \alpha_3$ are weighting coefficients. The intensity and gradient terms are

$$\mathcal{L}_{int} = \frac{1}{HW} \left\| \mathbf{Y}_f - \max(\mathbf{V}_{ir}, \mathbf{V}_{vi}) \right\|_1, \tag{14}$$

$$\mathcal{L}_{grad} = \frac{1}{HW} \left\| |\nabla \mathbf{Y}_f| - \max(|\nabla \mathbf{V}_{ir}|, |\nabla \mathbf{V}_{vi}|) \right\|_1, \tag{15}$$

where ∇ denotes the Sobel gradient operator. The color term keeps the chrominance of the visible source, penalizing deviation on the Cb and Cr channels only so that luminance stays free to follow the intensity and gradient terms:

$$\mathcal{L}_{color} = \frac{1}{HW} \sum_{c \in \{Cb, Cr\}} \left\| \mathbf{Y}_f^c - \mathbf{V}_{vi}^c \right\|_1. \tag{16}$$

Fidelity Loss. Two lightweight auxiliary decoders $\mathcal{D}_{ir}, \mathcal{D}_{vi}$ reconstruct the source frames from $\mathbf{F}_{frame}$:

$$\begin{aligned} \mathcal{L}_{fid} = \frac{1}{2} \Big(& \left\| \mathcal{D}_{ir}(\mathbf{F}_{frame}) - \mathbf{V}_{ir} \right\|_1 \\ & + \left\| \mathcal{D}_{vi}(\mathbf{F}_{frame}) - \mathbf{V}_{vi} \right\|_1 \Big). \end{aligned} \tag{17}$$

This term is used only during training and discarded at inference.

Temporal Consistency Loss. We penalize temporal variations that deviate from both sources:

$$\mathcal{L}_{tc} = \frac{1}{T-1} \sum_{t=1}^{T-1} \Big[\big(1 - \cos(\Delta \mathbf{Y}_f^t,\, \Delta \mathbf{V}_{vi}^t)\big) + \big(1 - \cos(\Delta \mathbf{Y}_f^t,\, \Delta \mathbf{V}_{ir}^t)\big) \Big], \tag{18}$$

where $\Delta \mathbf{Y}_f^t = \mathbf{Y}_f^{t+1} - \mathbf{Y}_f^t$ denotes the inter-frame difference, $\Delta \mathbf{V}_{vi}^t$ and $\Delta \mathbf{V}_{ir}^t$ are defined analogously, and $\cos(\cdot, \cdot)$ is cosine similarity. Complementing the tangent-space modeling in LEM, this pixel-domain term aligns the temporal variation direction of the fused video with both sources, suppressing flickering without optical flow.