MAMBAVF: STATE SPACE MODEL FOR EFFICIENT VIDEO FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video fusion aims to integrate complementary information from multiple source videos while preserving temporal consistency. Effective modeling of temporal dynamics is essential to this goal, yet existing methods incur substantial computational overhead from optical flow estimation and feature warping. In this paper, we present **MambaVF**, an efficient video fusion framework that uses state space model (SSM) to achieve temporal modeling without explicit motion estimation. First, by formulating video fusion as a sequential state update process, MambaVF captures long-range temporal dependencies with linear complexity, significantly reducing computation and memory costs. Second, the lightweight SSM-based fusion module eliminates conventional flow-guided alignment. Instead, it introduces a mutual state fusion module and a spatio-temporal bidirectional scanning mechanism to enable information aggregation across video streams. Experiments on multiple benchmarks confirm that MambaVF reaches state-of-the-art performance in different video fusion applications (multi-exposure, multi-focus, infrared-visible, medical), while reducing parameters by >90% and FLOPs by >80%, resulting in >50% shorter runtime. Code will be released.

1 INTRODUCTION

Video fusion aims to integrate complementary information from different sensors (*e.g.*, infrared and visible, or CT and MRI) or acquisition conditions (*e.g.*, different exposure times or focus settings) into a unified video sequence (Zhao et al., 2025). The enriched output video can significantly enhance video interpretation (Zhao et al., 2023b; Liu et al., 2022; Zhang, 2021). By mitigating sensing limitations and degradations in the individual inputs, fusion produces more informative videos for applications including computational imaging (Chen et al., 2025), autonomous driving (Prakash et al., 2021), robotic perception (Li et al., 2023a), and clinical diagnosis (Li et al., 2023c).

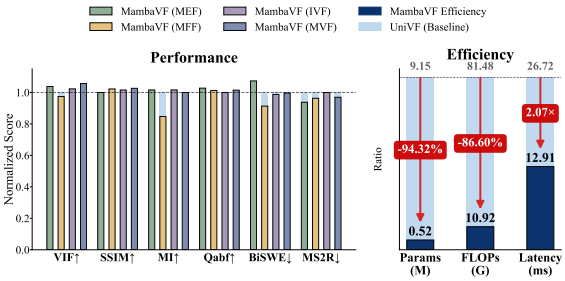

Figure 1: Compared with UniVF (Zhao et al., 2025), our MambaVF not only attains state-of-the-art performance on VF-Bench (Zhao et al., 2025), but requires only **5.7%** of the parameters and **13.4%** of the FLOPs, while achieving a **2.1×** speedup.

Existing video fusion methods typically require explicit optical flow estimation between consecutive frames, followed by feature warping, to aggregate information (Zhao et al., 2025). Such a flow-based approach faces two fundamental bottlenecks. (i) *Computational cost:* optical flow estimation and warping are expensive. As illustrated in Fig. 2 for UniVF (Zhao et al., 2025), they account for ≈80% of the total runtime and FLOPs, and ≈92% of all parameters. (ii) *Sensitivity to flow errors:* the quality of fusion is upper-bounded by the correctness of the estimated flow field. Occlusions, fast motion, or illumination changes that degrade the flow directly cause blur and ghosting artifacts in the fusion result. Conceptually, Transformers would make it possible to globally and robustly aggregate spatio-temporal information across frames via self-attention. But that would incur an even higher computational burden, due to the quadratic complexity of the attention mechanism. The high computational cost renders naive Transformer architectures impractical for video fusion,

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.

especially since applications like infrared-visible, multi-exposure and multi-focus fusion should often be deployed on-device, *i.e.*, on edge hardware and in near real-time.

Meanwhile, it has been shown that state space models (SSMs) such as the Mamba architecture (Gu & Dao, 2024) are often able to capture long-range dependencies almost as well as classical Transformers, but with linear computational complexity. Mamba's efficient latent-space updating mechanism is a natural fit for video sequences, a data type characterized by high temporal redundancy and long-term correlations (Li et al., 2024; Argaw et al., 2025). Of particular interest for our problem is that Mamba enables feature extraction in space and time without explicit, computationally heavy alignment and warping modules based on optical flow.

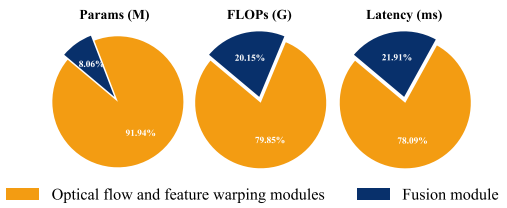

Figure 2: Parameter, FLOPs, and runtime contribution of optical flow and feature warping modules in UniVF (Zhao et al., 2025).

Motivated by this insight, we propose **MambaVF**, an efficient, SSM-driven video fusion framework. It departs from the usual practice of explicitly computing optical flow for frame-to-frame alignment. Instead, it leverages a *mutual state fusion* (MuFuse) module that captures the temporal dynamics of both input videos in a unified manner, enabling interaction between their hidden states. To ensure the model can comprehensively represent complex video content, we complement MuFuse with a *spatio-temporal bidirectional* (STB) scanning mechanism. Compared to standard 1D or 2D scanning, it performs a comprehensive traversal across spatial coordinates and frame indices. STB scanning improves spatio-temporal awareness in the latent representation, to ensure the model preserves the most informative structures from both inputs and outputs a high-fidelity fusion result.

In summary, the contributions of our work are:

- **MambaVF**, to our knowledge the first video fusion framework based on SSMs. By eliminating the need for optical flow computation and warping, it offers a new, highly efficient approach to video fusion.
- The **MuFuse** Mamba Encoder and associated **STB** scanning mechanism, which extend the standard Mamba block towards holistic video understanding across two (or more) different, complementary source streams.
- An extensive evaluation on VF-Bench, in which MambaVF matches the state-of-the-art performance of UniVF, while using only **5.7%** of the parameters and **13.4%** of the FLOPs, and running **2.1×** faster.

2 Related Work

Image and Video Fusion. Deep learning-based image fusion (Zhao et al., 2023b; Zhang, 2021) can be broadly divided into discriminative and generative paradigms. Discriminative methods (Xu et al., 2021; Zhao et al., 2022a; Li et al., 2023b; Zhao et al., 2022b) use CNNs (Prabhakar et al., 2017; Han et al., 2022; Gao et al., 2022; Bouzos et al., 2023; Guan et al., 2023; Deng et al., 2021) and Transformers (Qu et al., 2022; Guan et al., 2023) via model- or data-driven strategies (Xu et al., 2021; Zhao et al., 2022a; Li et al., 2023b; Zhao et al., 2022b; 2020; 2023b; Li & Wu, 2018). Generative approaches (Ma et al., 2019; Liu et al., 2022) treat fusion as distribution fitting, evolving from GANs (Ma et al., 2019; 2020) to Diffusion models (Chen et al., 2025; Zhao et al., 2023c), and recently to efficient one-step generation via rectified flow (Wang et al., 2025). To obtain more robust results, methods employ registration (Tang et al., 2025; Wang et al., 2022a; Jiang et al., 2022; Xu et al., 2022b) for misalignment and combined restoration for complex degradations (Cao et al., 2025b; Zhang et al., 2025). Concurrently, fusion is guided by downstream task feedback (Liu et al., 2025; Bai et al., 2025b; Zhong et al., 2025), cross-task synergy (Cheng et al., 2025; Xu et al., 2022a; Liang et al., 2022; Zhu et al., 2024b), and meta-learning (Bai et al., 2025c;a; Zhao et al., 2023a). Recently, vision-language priors (Zhao et al., 2024b) and text prompts (Zhang et al., 2024) have emerged for interactive, controllable fusion.

Then, to enhance the understanding of dynamics, image fusion has been extended to the video domain. Xie et al. (2024); Tang et al. (2026); Gong et al. (2025); Liu et al. (2026); Li et al. (2026c;b) mainly

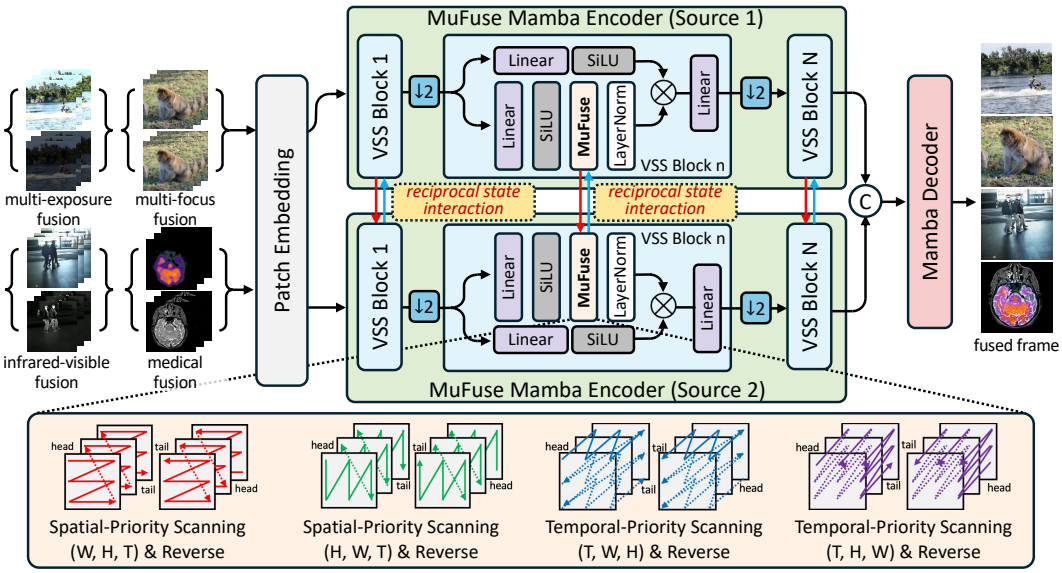

Figure 3: An overview of our **MambaVF** architecture. The MuFuse Mamba Encoder processes the embedded source features through VSS Blocks, whose MuFuse modules exchange complementary information via reciprocal state interaction. STB scanning supplies eight matching paths for the two sources: four forward orders with their full reversals. Tuples list axes from fastest to slowest. Spatial downsampling reduces encoding cost, and the decoder reconstructs the fused center frame.

focus on multi-modal scenarios. Zhao et al. (2025) propose UniVF, a unified video fusion framework that provides a comprehensive multi-frame learning paradigm for multiple fusion tasks, including multi-exposure, multi-focus, infrared-visible, and medical video fusion. This framework standardizes the learning objective, datasets, and evaluation protocols across these diverse scenarios.

State Space Models in Vision. State space models (Gu et al., 2021) provide an efficient basis for sequence modeling. In particular, Mamba (Gu & Dao, 2024) has shown strong performance in natural language processing by introducing a selective scan mechanism. Mamba demonstrates powerful capabilities in modeling long-range dependencies with linear computational complexity, bridging the gap between the efficiency of CNNs and the global receptive field of Transformers. Given the success of these efficient designs, many studies have adapted 1D SSMs to 2D visual tasks. Architectures like Vim (Zhu et al., 2024a) and VMamba (Liu et al., 2024a) use bidirectional or cross-scanning strategies to process images in visual recognition. In low-level vision, MambaIR (Guo et al., 2024) utilizes SSMs to restore pixel-level image details. Furthermore, recent works like VideoMamba (Li et al., 2024), Vivim (Yang et al., 2025), Vamba (Ren et al., 2025) and Video Mamba Suite (Chen et al., 2026) have extended this to the spatiotemporal domain, providing an efficient way to maintain temporal consistency without substantial computational overhead. Inspired by this, we aim to introduce the Mamba architecture to multi-source video fusion tasks, utilizing its efficient global modeling ability to handle cross-modal interactions and temporal evolution simultaneously.

3 MambaVF

In this section, we present MambaVF, an efficient state space model for unified video fusion. We first outline the framework, then describe the *Mutual State Fusion* (MuFuse) Mamba Encoder, *Spatio-Temporal Bidirectional* (STB) Scanning mechanism, and frame reconstruction.

3.1 Overview

Given synchronized source video windows $S_1, S_2 \in \mathbb{R}^{B\times T\times C\times H\times W}$, where B, T, C, H, W denote the batch size, number of frames, channels, height, and width, respectively, MambaVF generates a fused video sequence that integrates complementary information without explicit motion estimation.

The overall pipeline consists of three main stages, as illustrated in Fig. 3. First, the input videos are processed by a shared patch embedding layer to map raw pixels into a latent embedding space.

Second, a dual-stream MuFuse Mamba Encoder jointly models temporal dependencies and cross-source interactions, with spatial downsampling to reduce computation. Third, the two streams are concatenated and passed to a Mamba Decoder with PixelShuffle upsampling to reconstruct the central frame of each input window. This multiscale Mamba encoder-decoder aggregates spatio-temporal context through hidden states and reconstructs the fused frame without flow-guided alignment.

3.2 Feature Extraction and Interaction

Feature Embedding. MambaVF utilizes a dual-stream architecture to preserve the distinct characteristics of each source (*e.g.*, the thermal radiation in infrared and the texture details in visible videos). The two streams share a patch embedding layer:

$$E_S^1 = \text{Embedding}(S_1), \quad E_S^2 = \text{Embedding}(S_2), \tag{1}$$

producing a spatio-temporal feature volume for each source. The MuFuse Mamba Encoder processes these features through *Visual State Space* (VSS) Blocks, each containing a MuFuse Module that operates on sequences constructed by STB Scanning. Selective scanning scales linearly with the token count for fixed feature and state dimensions.

MuFuse Mamba Encoder. For a single source or modality, Mamba (Gu & Dao, 2024) scans a token sequence by updating a hidden state and computing contextual outputs as follows:

$$h_t = \bar{A}_t h_{t-1} + \bar{B}_t x_t, \quad y_t = C_t h_t + D x_t. \tag{2}$$

Here, $\bar{A}_t$ and $\bar{B}_t$ govern state propagation and input injection, while C_t maps the state to the output and D provides a direct input contribution. Mamba conditions B_t, C_t, and the discretization step size Δ_t on the current input, resulting in input-adaptive state updates.

Video fusion hinges on two equally important capabilities: extracting distinctive spatio-temporal cues from each source and exchanging complementary context between sources. The MuFuse Mamba Encoder therefore couples source-specific state modeling with reciprocal state interaction, so that each stream preserves its own contextual representation while benefiting from the other source's accumulated information. Within each VSS Block in the encoder, the two streams are projected into scan features and gating branches. The scan features follow matching STB paths and are processed by the MuFuse Module. The resulting outputs are restored, normalized, modulated by the corresponding source's SiLU gate, and projected back to the original feature width.

The MuFuse Module implements this interaction by combining source-specific state evolution with reciprocal state fusion. Along each pair of matching one-dimensional STB sequences (detailed below), each source independently accumulates context in its hidden state:

$$h_t^m = \bar{A}_t^m h_{t-1}^m + \bar{B}_t^m x_t^m, \quad m \in \{1, 2\}, \tag{3}$$

where t indexes scan positions and x_t^m is a projected source token. The states are coupled at the output through an input-dependent block projection:

$$\begin{bmatrix} y_t^1 \\ y_t^2 \end{bmatrix} = \underbrace{\begin{bmatrix} C_{1,t} & C_{2,t} \\ C_{3,t} & C_{4,t} \end{bmatrix}}_{\mathcal{C}_t\text{: reciprocal state fusion}} \begin{bmatrix} h_t^1 \\ h_t^2 \end{bmatrix} + \begin{bmatrix} D_1 x_t^1 \\ D_2 x_t^2 \end{bmatrix}. \tag{4}$$

The diagonal blocks preserve source-specific state readout, while the off-diagonal blocks enable reciprocal cross-source readout, allowing each output to incorporate the other source's state context. $C_{1,t}$ and $C_{3,t}$ are generated from x_t^1, whereas $C_{2,t}$ and $C_{4,t}$ are generated from x_t^2, using four distinct learned projections. Each state summarizes a scan history, so both streams receive contextual information from both sources while retaining separate within-block state dynamics. The interaction-enhanced outputs feed subsequent VSS Blocks.

The two entries in each column of $\mathcal{C}_t$ act on the same source state. A direct implementation nevertheless makes four scan calls, recomputing each source recurrence for self and cross outputs. We extend Mamba's low-level selective-scan operator through its C++ interface and CUDA forward/backward kernels to reuse each recurrence across both outputs, reducing four calls to two. This mathematically equivalent optimization reduces redundant computation while preserving numerical accuracy, as verified by FP32 forward and backward tests.

Finally, each VSS block is followed by $2\times$ spatial average pooling, which preserves temporal resolution while reducing the sequence length at the next stage to one quarter. This lowers the computational cost of context extraction and cross-source interaction.

Spatio-Temporal Bidirectional (STB) Scanning. STB Scanning organizes video features into complementary spatial-priority and temporal-priority sequences for the MuFuse Module. Each of four forward orders is paired with its full reversal, yielding eight directions as illustrated in Fig. 3.

Let $x^m \in \mathbb{R}^{d\times T\times H_\ell\times W_\ell}$ denote a source feature volume at encoder stage ℓ, where d is the feature width. Axis tuples are ordered from fastest to slowest. The paths are organized as follows:

1. *Spatial-Priority Scanning*: (W, H, T) traverses each frame row by row, whereas (H, W, T) traverses it column by column, before advancing in time. These orders make horizontal or vertical neighbors adjacent in the sequence, supporting intra-frame texture and structure aggregation.
2. *Temporal-Priority Scanning*: (T, W, H) and (T, H, W) visit all frames at a fixed spatial location before advancing to the next location in row-major or column-major order. They reduce the sequence distance between adjacent frames at the same coordinate from $H_\ell W_\ell$ in the spatial-priority orders to one step, capturing temporal variation without explicitly tracking motion trajectories.
3. *Symmetrical Bidirectional Traversal*: Each forward path is paired with the reverse of its entire sequence. The two sources follow matching paths, so their states summarize corresponding scan histories for reciprocal state interaction.

After the multi-path scanning, the eight sequences are mapped back to their original spatio-temporal dimensions and integrated to form the enhanced feature representation. Existing Mamba variants explore adaptive spatial ordering and multiple spatiotemporal scanning routes (Li et al., 2026a; Feng et al., 2025), while interaction-oriented designs organize scans across tasks or interleave modalities along spatial directions (Cao et al., 2025a; Wang et al., 2026). Our MuFuse Module combines eight spatiotemporal scanning paths with complementary spatial and temporal priorities, applying the same traversal order to both sources along each path. The distinction lies in coupling shared spatiotemporal traversal with state interaction. Each scanning path models context within each source while exchanging information between the two sources.

3.3 Fusion and Reconstruction

The MuFuse Mamba Encoder $\mathcal{E}_{\text{MuFuse}}$ produces two feature volumes F_1, F_2, which are merged by channel-wise concatenation:

$$F_1, F_2 = \mathcal{E}_{\text{MuFuse}}(E_S^1, E_S^2),\ F_{\text{fused}} = \text{Concat}([F_1, F_2]). \tag{5}$$

The fused features F_{fused} are subsequently processed by the Mamba Decoder, which applies three stages of $2\times$ PixelShuffle upsampling with spatial Mamba refinement. Accordingly, it uses four-way spatial scanning for center-frame reconstruction. This design choice concentrates the state-space modeling on intra-frame structural refinement, ensuring high-fidelity spatial reconstruction for the target frame. Finally, 2D residual blocks (He et al., 2016) are lastly used to restore fine-grained textures and project the high-dimensional embeddings back to the pixel domain.

4 Experiments

MambaVF is evaluated on VF-Bench (Zhao et al., 2025) across four diverse scenarios: *multi-exposure fusion* (MEF), *multi-focus fusion* (MFF), *infrared-visible fusion* (IVF), and *medical video fusion* (MVF). We report experimental setup, fusion quality, computational efficiency, and ablation studies.

4.1 Experimental Configuration

Loss Function. Following the training setup of UniVF (Zhao et al., 2025), MambaVF is trained using a multi-term loss function that balances spatial, structural, and temporal fidelity:

$$\mathcal{L} = \mathcal{L}_{\text{spatial}} + \alpha_1 \mathcal{L}_{\text{grad}} + \alpha_2 \mathcal{L}_{\text{temp}}. \tag{6}$$

The individual loss components are as follows:

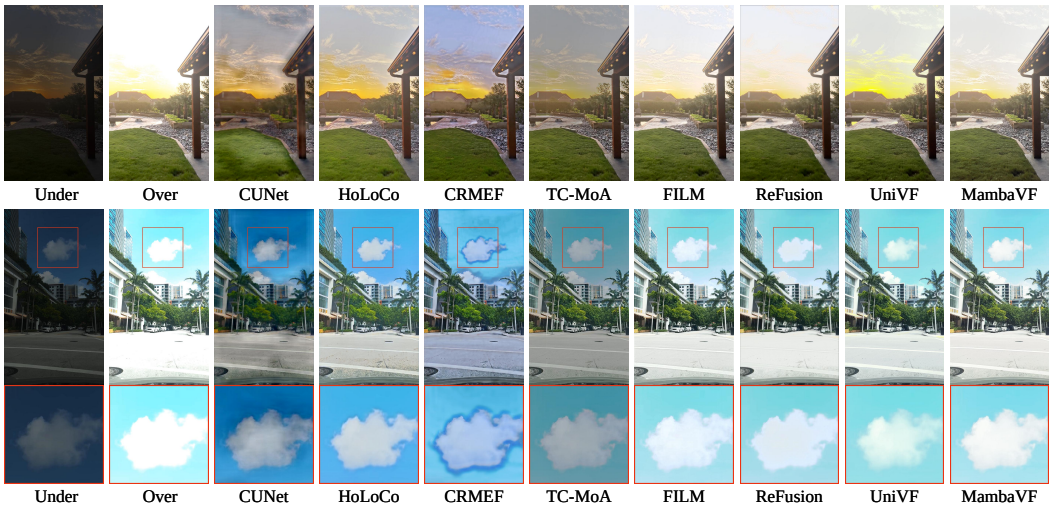

Figure 4: Visual comparison of fused results for multi-exposure video fusion.

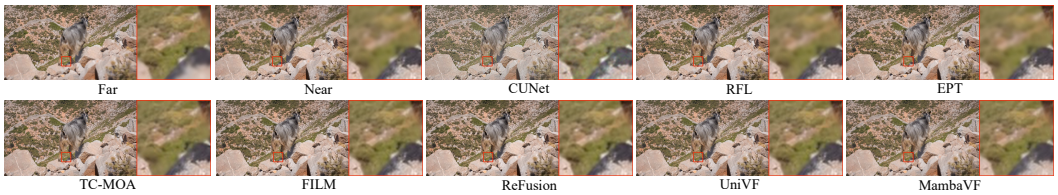

Figure 5: Visual comparison of fused results for multi-focus video fusion.

(a) *Spatial Similarity* ($\mathcal{L}_{\text{spatial}}$): This term supervises the photometric distribution of the fused output. For IVF and MVF, we maximize intensity saliency: $\mathcal{L}_{\text{spatial}} = \frac{1}{HW}\|I_t^F - \max(I_t^1, I_t^2)\|_1$. For MEF, it combines a mean intensity term $\mathcal{L}_{int} = \frac{1}{HW}\|I_t^F - \text{mean}(I_t^1, I_t^2)\|_1$ and the MEF-SSIM loss (Ma et al., 2015). For MFF, $\mathcal{L}_{\text{spatial}}$ is equivalent to $\mathcal{L}_{int} = \frac{1}{HW}\|I_t^F - \text{mean}(I_t^1, I_t^2)\|_1$.

(b) *Gradient Preservation* ($\mathcal{L}_{\text{grad}}$): To sharpen fused textures, we utilize a Sobel-based gradient loss: $\mathcal{L}_{\text{grad}} = \frac{1}{HW}\||\nabla I_t^F| - \max(|\nabla I_t^1|, |\nabla I_t^2|)\|_1$, encouraging the retention of high-frequency structural details from both sources.

(c) *Temporal Consistency* ($\mathcal{L}_{\text{temp}}$): This term penalizes abrupt inter-frame intensity fluctuations, ensuring smooth photometric transitions without explicit motion compensation. For a detailed derivation of these objectives and settings, please refer to Zhao et al. (2025).

Implementation Details. The MambaVF framework is implemented using PyTorch and trained on 4×NVIDIA GH200 GPUs. We utilize the Adam optimizer with an initial learning rate of 1×10^{-4}, which is gradually reduced through an exponential decay strategy over 20k iterations. The training process uses a batch size of 32 with gradient accumulation. The MuFuse Mamba Encoder has three VSS Blocks with 32 feature channels per source.

Evaluation Protocol. Following the evaluation protocol in VF-Bench (Zhao et al., 2025), our evaluation encompasses both spatial and temporal dimensions. For spatial quality, we employ *Visual Information Fidelity* (VIF), *Structural Similarity* (SSIM), *Mutual Information* (MI), and $Q^{AB/F}$, assessing perceptual fidelity, structural similarity, shared information, and edge preservation, respectively. For temporal stability, we adopt *Bi-Directional Self-Warping Error* (BiSWE) and *Motion Smoothness with Dual Reference Videos* (MS2R) to measure inter-frame warping consistency and agreement with source-video flow changes, respectively. Higher values in spatial metrics and lower values in temporal metrics indicate superior fusion performance.

4.2 Video Fusion Experiments

We compare MambaVF with SOTA methods on four video fusion tasks to demonstrate its superior performance. Qualitative video comparison results are provided in the supplementary material.

Multi-Exposure Video Fusion. The MEF task requires the model to synthesize high-dynamic-range content from sequences with varying exposure levels. We compare MambaVF against SOTA

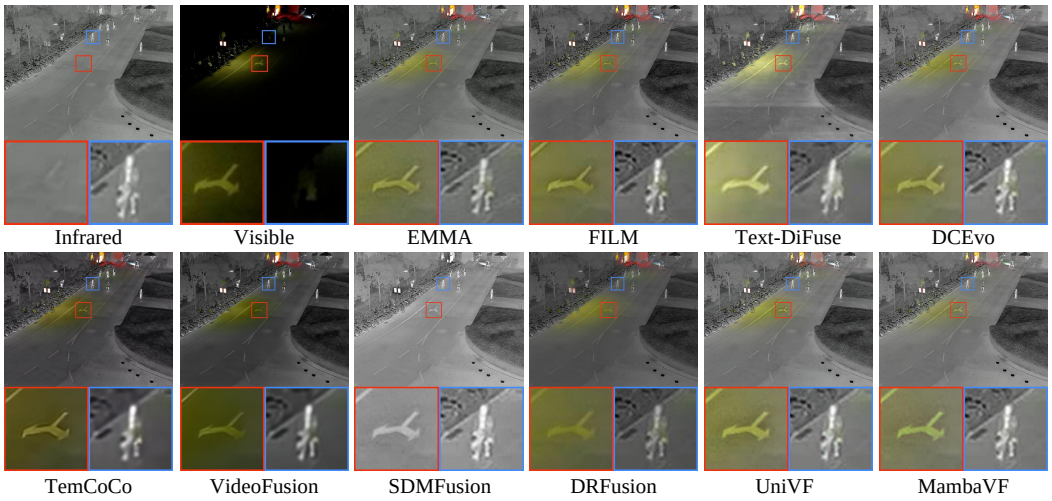

Figure 6: Visual comparison of fused results for infrared-visible video fusion.

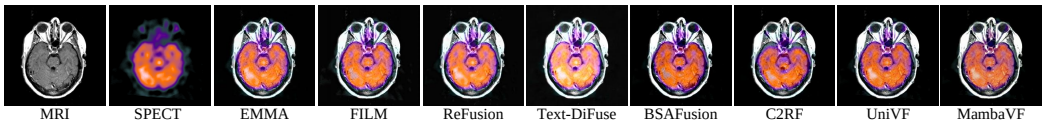

Figure 7: Visual comparison of fused results for MVF.

methods including CUNet (Deng & Dragotti, 2020), HoLoCo (Liu et al., 2023), CRMEF (Liu et al., 2024b), TC-MoA (Zhu et al., 2024b), FILM (Zhao et al., 2024b), ReFusion (Bai et al., 2025c), and UniVF (Zhao et al., 2025). As illustrated in Fig. 4, MambaVF demonstrates a superior ability to recover details in both extremely dark and saturated regions, modeling the temporal evolution of luminance implicitly. This leads to more natural transitions of light and shadow, achieving higher VIF scores and significantly lower MS2R values in Tab. 1, proving that MambaVF can maintain high-fidelity textures while ensuring temporal alignment in dynamic lighting conditions.

Multi-Focus Video Fusion. In MFF, the objective is to integrate the focused regions from different focal planes into an all-in-focus video. We evaluate MambaVF against CUNet (Deng & Dragotti, 2020), RFL (Wang et al., 2022b), EPT (Wang et al., 2023), TC-MoA (Zhu et al., 2024b), FILM (Zhao et al., 2024b), ReFusion (Bai et al., 2025c), and UniVF (Zhao et al., 2025). The VF-Bench results in Fig. 5 and Tab. 1 show that MambaVF reduces halo artifacts near focus boundaries by capturing spatial and temporal sharpness cues. Our method accurately identifies the transition zones between focused and defocused areas, preserving sharp edges across the entire sequence. MambaVF outperforms baseline models, particularly in scenes with complex depth-of-field changes where traditional image-based methods struggle to maintain boundary consistency.

Infrared-Visible Video Fusion. We compare our model with EMMA (Zhao et al., 2024a), FILM (Zhao et al., 2024b), Text-DiFuse (Zhang et al., 2024), DCEvo (Liu et al., 2025), TemCoCo (Gong et al., 2025), VideoFusion (Tang et al., 2026), SDMFusion (Liu et al., 2026), DRFusion (Li et al., 2026c), and UniVF (Zhao et al., 2025). MambaVF excels at preserving the high-contrast thermal signatures of infrared sources while retaining the rich environmental details of visible light. Qualitatively, our fused videos exhibit more prominent targets and clearer background structures in Fig. 6. Quantitatively, in Tab. 2, MambaVF achieves leading performance in VIF, SSIM and BiSWE, indicating that it successfully integrates the maximum amount of complementary information from both modalities while preserving video fluidity.

Medical Video Fusion. For the MVF task, we fuse MRI images with CT, PET, or SPECT images. We compare MambaVF with EMMA (Zhao et al., 2024a), FILM (Zhao et al., 2024b), ReFusion (Bai et al., 2025c), Text-DiFuse (Zhang et al., 2024), BSAFusion (Li et al., 2025), C2RF (Tang et al., 2025), and UniVF (Zhao et al., 2025). MambaVF demonstrates exceptional performance in maintaining modality information. It preserves fine-grained tissue textures of MRI images while successfully superimposing the functional intensity distributions from PET or SPECT. The results in Tab. 3

Table 1: Quantitative evaluation results for the low-resolution MEF and MFF tasks. The red and blue highlights indicate the best and second-best scores.

	VF-Bench Multi-Exposure Fusion Branch (540p)							VF-Bench Multi-Focus Fusion Branch (480p)					
	VIF↑	SSIM↑	MI↑	Qabf↑	BiSWE↓	MS2R↓		VIF↑	SSIM↑	MI↑	Qabf↑	BiSWE↓	MS2R↓
CUNet	0.50	0.85	1.85	0.39	7.55	0.20	CUNet	0.53	0.86	3.52	0.68	10.23	0.42
HoLoCo	0.50	0.86	2.56	0.42	8.22	0.19	RFL	0.77	0.90	6.31	0.78	8.46	0.28
CRMEF	0.62	0.94	2.60	0.63	8.72	0.19	EPT	0.76	0.90	6.33	0.78	8.50	0.29
TC-MoA	0.74	0.99	2.93	0.72	7.82	0.16	TC-MoA	0.75	0.90	5.27	0.77	8.39	0.28
FILM	0.77	0.99	4.35	0.72	8.28	0.17	FILM	0.75	0.89	5.06	0.78	8.61	0.33
ReFus	0.74	0.97	3.81	0.72	7.63	0.16	ReFus	0.73	0.90	4.93	0.77	8.00	0.32
UniVF	0.79	0.99	4.38	0.73	6.96	0.16	UniVF	0.77	0.90	6.34	0.79	8.29	0.27
MambaVF	0.82	0.99	4.45	0.75	7.47	0.15	MambaVF	0.75	0.92	5.37	0.80	7.57	0.26

Table 2: Quantitative evaluation for IVF.

	VF-Bench Infrared-Visible Video Fusion Branch					
	VIF ↑	SSIM ↑	MI ↑	Qabf ↑	BiSWE ↓	MS2R ↓
EMMA	0.37	0.63	2.01	0.58	4.79	0.37
FILM	0.40	0.63	2.05	0.64	4.78	0.37
Text-D	0.30	0.60	1.64	0.39	10.63	0.40
DCEvo	0.43	0.64	2.44	0.66	4.57	0.37
TemCoCo	0.29	0.60	1.77	0.41	4.66	0.46
VideoFusion	0.27	0.60	1.83	0.37	3.00	0.57
SDMFusion	0.43	0.61	1.94	0.65	4.49	0.37
DRFusion	0.35	0.64	1.98	0.53	4.90	0.38
UniVF	0.44	0.64	2.47	0.68	3.94	0.35
MambaVF	0.45	0.65	2.51	0.68	3.89	0.35

Table 3: Quantitative evaluation for MVF.

	VF-Bench Medical Video Fusion Branch					
	VIF ↑	SSIM ↑	MI ↑	Qabf ↑	BiSWE ↓	MS2R ↓
EMMA	0.29	0.68	1.73	0.60	30.00	1.98
FILM	0.33	0.36	1.83	0.67	32.04	1.59
ReFus	0.31	0.32	1.74	0.67	32.85	1.74
Text-D	0.24	0.21	1.58	0.52	34.09	1.96
BSAF	0.28	0.63	1.69	0.58	34.73	1.66
C2RF	0.30	0.73	1.75	0.59	32.67	2.06
UniVF	0.35	0.76	2.00	0.68	29.61	1.30
MambaVF	0.37	0.78	2.00	0.69	29.51	1.26

and Fig. 7 show that our model achieves the highest structural similarity scores and interframe similarity, effectively capturing complex spatial relationships within medical volumes.

4.3 Efficiency Analysis

A cornerstone of MambaVF is its superior computational efficiency, which is particularly vital for real-time video processing. We compare MambaVF with all video fusion competitors, including the *unified* video fusion framework UniVF (Zhao et al., 2025) and four *IVF-specific* video fusion methods: TemCoCo (Gong et al., 2025), VideoFusion (Tang et al., 2026), SDMFusion (Liu et al., 2026), and DRFusion (Li et al., 2026c).

As detailed in Fig. 1 and Tab. 5, MambaVF requires only **5.7%** of the parameters and **13.4%** of the FLOPs compared with the unified fusion baseline UniVF, achieving a **2.1**× speedup. This drastic reduction in computational footprint stems primarily from our flow-free design, which bypasses the heavy resource requirements of motion estimation and feature warping modules.

Compared with the IVF-specific methods, MambaVF has the lowest reported parameter and FLOP counts while supporting all four fusion tasks. Its inference latency is lower than those of VideoFusion, SDMFusion, and DRFusion, and very close to that of TemCoCo. Thus, MambaVF combines broader task coverage with low reported complexity and competitive latency.

4.4 Ablation Studies

To further investigate the impact of our design choices, we conduct several controlled experiments on the MEF task in Tab. 4.

Architectural Components. We evaluate STB Scanning and the MuFuse Module against three restricted baselines:

(1) Exp. I: 2D spatial scan. Replace STB scanning with a standard 4-way 2D spatial scan;
(2) Exp. II: 1D temporal scan. Replace STB scanning with a simplified 1D temporal-only scan;
(3) Exp. III: w/o state interaction. Remove the cross-source terms $C_{2,t}h_t^2$ and $C_{3,t}h_t^1$ in Eq. (4), while retaining each source's independent Mamba scan and the full STB Scanning schedule.

The quantitative results show that 2D spatial scanning cannot model inter-frame dependencies, leading to higher BiSWE and visible flickering, while 1D temporal scanning captures motion but sacrifices fine-grained spatial structures. In contrast, our 8-way STB scanning jointly captures pixel evolution and intra-frame geometry from multiple perspectives, achieving a better spatiotemporal balance. Removing reciprocal state interaction consistently degrades performance, demonstrating that

Table 4: Results of the ablation experiments, with the best-performing values highlighted in red.

Descriptions	Configurations			Metrics					
	Scan type	3D Decoder	Multi-frame	VIF ↑	SSIM ↑	MI ↑	Qabf ↑	BiSWE ↓	MS2R ↓
Exp. I: 2D spatial scan	2D	✗	✓	0.74	0.98	3.83	0.70	8.02	0.20
Exp. II: 1D temporal scan	1D	✗	✓	0.62	0.95	2.97	0.58	8.35	0.17
Exp. III: w/o state interaction	3D	✗	✓	0.79	0.99	4.16	0.73	7.87	0.16
Exp. IV: Single-frame input	2D	✗	✗	0.68	0.98	3.49	0.70	7.94	0.18
Exp. V: 3D Decoder	3D	✓	✓	0.79	0.99	3.86	0.73	7.68	0.16
MambaVF (Ours)	3D	✗	✓	0.82	0.99	4.45	0.75	7.47	0.15

Table 5: Efficiency comparison among video fusion methods[1]. Red and blue backgrounds mark the lowest and second-lowest values.

Methods	Task	Params (M)	FLOPs (G)	Latency (ms)
TemCoCo	IVF	19.18	158.54	12.70
VideoFusion	IVF	6.74	85.69	18.93
SDMFusion	IVF	835.86	244.99	94.35
DRFusion	IVF	86.62	35299.93	1236.15
UniVF	Unified	9.15	81.48	26.72
MambaVF	Unified	0.52	10.92	12.91

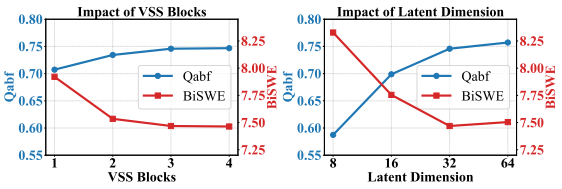

Figure 8: Hyperparameter analysis of VSS blocks and latent dimension.

exchanging source-specific state information improves both information preservation and temporal consistency beyond independent source modeling.

Temporal Context and Decoder Refinement. We further examine temporal context and reconstruction through two variants:

(4) Exp. IV: Single-frame input. Use only the target frame from each source, removing neighboring-frame context and degrading MambaVF into a frame-by-frame image fusion model.
(5) Exp. V: 3D Decoder ResBlocks. Replace the decoder's 2D residual blocks with their 3D counterparts to assess additional spatio-temporal processing during reconstruction.

The transition from multi-frame to single-frame input (Exp. IV) caused a substantial drop in overall fusion quality, underscoring the necessity of temporal context for maintaining coherence in video fusion. Regarding the decoder architecture (Exp. V), while 3D Residual Blocks aim to capture deeper spatio-temporal features, they actually yielded a slight performance decrease and introduced significant computational overhead. Adhering to the principle of parsimony, we retain the 2D Residual Blocks, as they provide superior texture refinement with significantly lower resource consumption.

Impact of VSS Blocks and Hidden State Dimension. We analyze the sensitivity to the number of VSS blocks and the latent dimension in Fig. 8. Increasing the depth of the Mamba encoder improves the fusion performance but introduces additional latency. We found that a depth of 3 blocks provides an optimal trade-off for most real-time scenarios. Furthermore, increasing the latent dimension beyond 32 does not lead to significant performance gains, but it does increase computational cost. Therefore, we choose 3 VSS blocks with an embedding dimension of 32 as our configuration.

5 Conclusion

We introduced MambaVF, an efficient and flow-free framework for multi-source video fusion based on state space models. By shifting away from the traditional estimate-and-warp paradigm, we address the long-standing bottlenecks of computational redundancy and flow-dependency in video fusion. In MambaVF, the MuFuse Mamba Encoder couples source-specific context extraction with reciprocal state interaction, while eight-direction STB Scanning aggregates complementary spatial and temporal information without explicit motion estimation or feature warping. Experimental results on VF-Bench show that MambaVF not only achieves SOTA performance across four fusion tasks but also offers a significant reduction in parameters, FLOPs and runtime. We believe MambaVF provides a new, high-performance baseline for the deployment of fusion algorithms on resource-constrained platforms and offers a promising direction for future research in efficient spatio-temporal modeling and fusion.

[1]All compared methods are video fusion methods; image fusion methods are not included.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the grammar, clarity, and readability of text written by the authors. Generative AI was not used to generate research ideas, formulate hypotheses or claims, design the methodology or experiments, analyze data, or interpret results. The authors reviewed all AI-assisted edits and take full responsibility for the final content of this work.

REFERENCES

Dawit Mureja Argaw, Xian Liu, Joon Son Chung, Ming-Yu Liu, and Fitsum Reda. Mambavideo for discrete video tokenization with channel-split quantization. *arXiv preprint arXiv:2507.04559*, 2025.

Haowen Bai, Jiangshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao Feng, and Shuang Xu. Task-driven image fusion with learnable fusion loss. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7457–7468, 2025a.

Haowen Bai, Zixiang Zhao, Jiangshe Zhang, Baisong Jiang, Lilun Deng, Yukun Cui, Shuang Xu, and Chunxia Zhang. Deep unfolding multi-modal image fusion network via attribution analysis. *IEEE Trans. Circuit Syst. Video Technol.*, 35(4):3498–3511, 2025b.

Haowen Bai, Zixiang Zhao, Jiangshe Zhang, Yichen Wu, Lilun Deng, Yukun Cui, Baisong Jiang, and Shuang Xu. Refusion: Learning image fusion from reconstruction with learnable loss via meta-learning. *Int. J. Comput. Vis.*, 133(5):2547–2567, 2025c.

Odysseas Bouzos, Ioannis Andreadis, and Nikolaos Mitianoudis. A convolutional neural network-based conditional random field model for structured multi-focus image fusion robust to noise. *IEEE Trans. Image Process.*, 32:2915–2930, 2023.

Mang Cao, Sanping Zhou, Yizhe Li, Ye Deng, Wenli Huang, and Le Wang. Enhancing mamba decoder with bidirectional interaction in multi-task dense prediction. In *Int. Conf. Comput. Vis.*, pp. 1–10. IEEE, 2025a.

Zihan Cao, Yu Zhong, Ziqi Wang, and Liang-Jian Deng. Mmaif: Multi-task and multi-degradation all-in-one for image fusion with language guidance. In *Int. Conf. Comput. Vis.*, pp. 11744–11754, October 2025b.

Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Jiahao Wang, Zhe Chen, Zhiqi Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding. *Int. J. Comput. Vis.*, 134(1):20, 2026.

Zixuan Chen, Yujin Wang, Xin Cai, Zhiyuan You, Zheming Lu, Fan Zhang, Shi Guo, and Tianfan Xue. Ultrafusion: Ultra high dynamic imaging using exposure fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 16111–16121, June 2025.

Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Zeyang Zhang, Sara Atito Ali, Muhammad Awais, and Josef Kittler. One model for ALL: low-level task interaction is a key to task-agnostic image fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 28102–28112, 2025.

Xin Deng and Pier Luigi Dragotti. Deep convolutional neural network for multi-modal image restoration and fusion. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3333–3348, 2020.

Xin Deng, Yutong Zhang, Mai Xu, Shuhang Gu, and Yiping Duan. Deep coupled feedback network for joint exposure fusion and image super-resolution. *IEEE Trans. Image Process.*, 30:3098–3112, 2021.

Runyang Feng, Hyung Jin Chang, Tze Ho Elden Tse, Boeun Kim, Yi Chang, and Yixing Gao. High-resolution spatiotemporal modeling with global-local state space models for video-based human pose estimation. In *Int. Conf. Comput. Vis.*, pp. 1–10. IEEE, 2025.

Fangyuan Gao, Xin Deng, Mai Xu, Jingyi Xu, and Pier Luigi Dragotti. Multi-modal convolutional dictionary learning. *IEEE Trans. Image Process.*, 31:1325–1339, 2022.

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

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conf. on Language Modeling*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Yuanshen Guan, Ruikang Xu, Mingde Yao, Lizhi Wang, and Zhiwei Xiong. Mutual-guided dynamic network for image fusion. In *ACM Int. Conf. Multimedia*, pp. 1779–1788, 2023.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. Mambair: A simple baseline for image restoration with state-space model. In *Eur. Conf. Comput. Vis.*, pp. 222–241. Springer, 2024.

Dong Han, Liang Li, Xiaojie Guo, and Jiayi Ma. Multi-exposure image fusion via deep perceptual enhancement. *Inf. Fusion*, 79:248–262, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 770–778, 2016.

Zhiying Jiang, Zengxi Zhang, Xin Fan, and Risheng Liu. Towards all weather and unobstructed multi-spectral image stitching: Algorithm and benchmark. In *ACM Int. Conf. Multimedia*, pp. 3783–3791, 2022.

Hao Li, Yizhi Zhang, Junzhe Zhu, Shaoxiong Wang, Michelle A Lee, Huazhe Xu, Edward Adelson, Li Fei-Fei, Ruohan Gao, and Jiajun Wu. See, hear, and feel: Smart sensory fusion for robotic manipulation. In *Conf. on Robot Learning*, pp. 1368–1378, 2023a.

Huafeng Li, Dayong Su, Qing Cai, and Yafei Zhang. Bsafusion: A bidirectional stepwise feature alignment network for unaligned medical image fusion. In *AAAI*, pp. 4725–4733, 2025.

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

Hui Li, Tianyang Xu, Xiaojun Wu, Jiwen Lu, and Josef Kittler. Lrrnet: A novel representation learning guided fusion network for infrared and visible images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):11040–11052, 2023b.

Jiawei Li, Jinyuan Liu, Shihua Zhou, Qiang Zhang, and Nikola K. Kasabov. Gesenet: A general semantic-guided network with couple mask ensemble for medical image fusion. *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1–14, 2023c.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *Eur. Conf. Comput. Vis.*, pp. 237–255. Springer, 2024.

Tanzhe Li, Caoshuo Li, Jiayi Lyu, Hongjuan Pei, Baochang Zhang, Taisong Jin, and Rongrong Ji. Damamba: Vision state space model with dynamic adaptive scan. *Adv. Neural Inform. Process. Syst.*, 38:39777–39799, 2026a.

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*, 2026b.

Xingyuan Li, Haoyuan Xu, Shulin Li, Xiang Chen, Zhiying Jiang, and Jinyuan Liu. Drfusion: Drift-resilient temporally consistent infrared-visible video fusion. In *Int. Conf. Mach. Learn.*, 2026c.

Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *Eur. Conf. Comput. Vis.*, pp. 719–735, 2022.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5792–5801, 2022.

Jinyuan Liu, Guanyao Wu, Junsheng Luan, Zhiying Jiang, Risheng Liu, and Xin Fan. Holoco: Holistic and local contrastive learning network for multi-exposure image fusion. *Inf. Fusion*, 95: 237–249, 2023.

Jinyuan Liu, Bowei Zhang, Qingyun Mei, Xingyuan Li, Yang Zou, Zhiying Jiang, Long Ma, Risheng Liu, and Xin Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2226–2235, 2025.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 14305–14314, June 2026.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Adv. Neural Inform. Process. Syst.*, 37: 103031–103063, 2024a.

Zhu Liu, Jinyuan Liu, Guanyao Wu, Zihang Chen, Xin Fan, and Risheng Liu. Searching a compact architecture for robust multi-exposure image fusion. *IEEE Trans. Circuit Syst. Video Technol.*, 34 (7):6224–6237, 2024b.

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

Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping (Steven) Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans. Image Process.*, 29:4980–4995, 2020.

Kede Ma, Kai Zeng, and Zhou Wang. Perceptual quality assessment for multi-exposure image fusion. *IEEE Trans. Image Process.*, 24(11):3345–3356, 2015.

K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Int. Conf. Comput. Vis.*, pp. 4724–4732, 2017.

Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7077–7087, 2021.

Linhao Qu, Shaolei Liu, Manning Wang, and Zhijian Song. Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning. In *AAAI*, pp. 2126–2134, 2022.

Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. Vamba: Understanding hour-long videos with hybrid mamba-transformers. In *Int. Conf. Comput. Vis.*, pp. 21197–21208, October 2025.

Linfeng Tang, Qinglong Yan, Leyuan Fang Xinyu Xiang, and Jiayi Ma. C2rf: Bridging multi-modal image registration and fusion via commonality mining and contrastive learning. *Int. J. Comput. Vis.*, 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 19559–19569, 2026.

Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *IJCAI*, pp. 3508–3515, 2022a.

Yingying Wang, Xuanhua He, Chen Wu, Jialing Huang, Suiyun Zhang, Rui Liu, Xinghao Ding, and Haoxuan Che. Mmmamba: A versatile cross-modal in context fusion framework for pan-sharpening and zero-shot image enhancement. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *AAAI*, pp. 18656–18664. AAAI Press, 2026.

Zeyu Wang, Xiongfei Li, Haoran Duan, and Xiaoli Zhang. A self-supervised residual feature learning model for multifocus image fusion. *IEEE Trans. Image Process.*, 31:4527–4542, 2022b.

Zeyu Wang, Xiongfei Li, Libo Zhao, Haoran Duan, Shidong Wang, Hao Liu, and Xiaoli Zhang. When multi-focus image fusion networks meet traditional edge-preservation technology. *Int. J. Comput. Vis.*, 131(10):2529–2552, 2023.

Zirui Wang, Jiayi Zhang, Tianwei Guan, Yuhan Zhou, Xingyuan Li, Minjing Dong, and Jinyuan Liu. Efficient rectified flow for image fusion. In *Adv. Neural Inform. Process. Syst.*, 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 Trans. Multimedia*, 2024.

Han Xu, Jiayi Ma, Junjun Jiang, Xiaojie Guo, and Haibin Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):502–518, 2022a.

Han Xu, Jiayi Ma, Jiteng Yuan, Zhuliang Le, and Wei Liu. Rfnet: Unsupervised network for mutually reinforcing multi-modal image registration and fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 19647–19656, 2022b.

Shuang Xu, Jiangshe Zhang, Zixiang Zhao, Kai Sun, Junmin Liu, and Chunxia Zhang. Deep gradient projection networks for pan-sharpening. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1366–1375, 2021.

Yijun Yang, Zhaohu Xing, Lequan Yu, Huazhu Fu, Chunwang Huang, and Lei Zhu. Vivim: A video vision mamba for ultrasound video segmentation. *IEEE Trans. Circuit Syst. Video Technol.*, 35 (10):10293–10304, 2025.

Hao Zhang, Lei Cao, and Jiayi Ma. Text-difuse: An interactive multi-modal image fusion framework based on text-modulated diffusion model. In *Adv. Neural Inform. Process. Syst.*, 2024.

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

Xingchen Zhang. Deep learning-based multi-focus image fusion: A survey and a comparative study. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):4819–4838, 2021.

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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 13955–13965. IEEE, 2023a.

Zixiang Zhao, Shuang Xu, Chunxia Zhang, Junmin Liu, Jiangshe Zhang, and Pengfei Li. DIDFuse: Deep image decomposition for infrared and visible image fusion. In *IJCAI*, pp. 970–976, 2020.

Zixiang Zhao, Shuang Xu, Jiangshe Zhang, Chengyang Liang, Chunxia Zhang, and Junmin Liu. Efficient and model-based infrared and visible image fusion via algorithm unrolling. *IEEE Trans. Circuit Syst. Video Technol.*, 32(3):1186–1196, 2022a.

Zixiang Zhao, Jiangshe Zhang, Shuang Xu, Zudi Lin, and Hanspeter Pfister. Discrete cosine transform network for guided depth map super-resolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5687–5697, 2022b.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5906–5916, 2023b.

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 *Int. Conf. Comput. Vis.*, pp. 8082–8093, 2023c.

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Kai Zhang, Shuang Xu, Dongdong Chen, Radu Timofte, and Luc Van Gool. Equivariant multi-modality image fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 25912–25921. IEEE, 2024a.

Zixiang Zhao, Lilun Deng, Haowen Bai, Yukun Cui, Zhipeng Zhang, Yulun Zhang, Haotong Qin, Dongdong Chen, Jiangshe Zhang, Peng Wang, and Luc Van Gool. Image fusion via vision-language model. In *Int. Conf. Mach. Learn.*, 2024b.

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 *Adv. Neural Inform. Process. Syst.*, 2025.

Haifeng Zhong, Fan Tang, Zhuo Chen, Hyung Jin Chang, and Yixing Gao. Amdanet: Attention-driven multi-perspective discrepancy alignment for rgb-infrared image fusion and segmentation. In *Int. Conf. Comput. Vis.*, pp. 10645–10655, October 2025.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024a.

Pengfei Zhu, Yang Sun, Bing Cao, and Qinghua Hu. Task-customized mixture of adapters for general image fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7099–7108. IEEE, 2024b.

A Additional Qualitative Fusion Comparison Results

Additional visual evidence is provided in the figures below to further demonstrate the performance of our framework:

- Further visual results for the *Multi-exposure Video Fusion* task are presented in Fig. 9.
- Further visual results for the *Multi-focus Video Fusion* task are presented in Fig. 10.
- Further visual results for the *Infrared-Visible Video Fusion* task are presented in Fig. 11.
- Further visual results for the *Medical Video Fusion* task are presented in Fig. 12.

These visual evaluations corroborate our primary findings: MambaVF faithfully recovers subtle structures and intricate textures from source inputs while successfully merging complementary features into high-quality, information-rich fused frames. Furthermore, the video sequences included in the supplementary materials highlight the robust temporal stability of our method, characterized by a significant reduction in inter-frame jitter and motion artifacts.

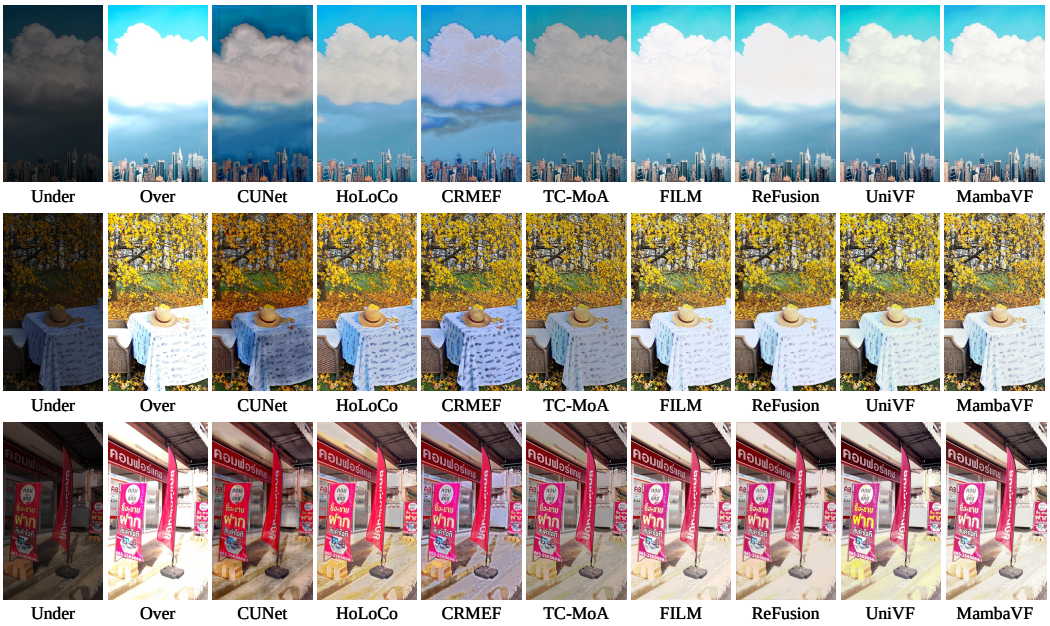

Figure 9: Visual comparison of fused results for multi-exposure video fusion.

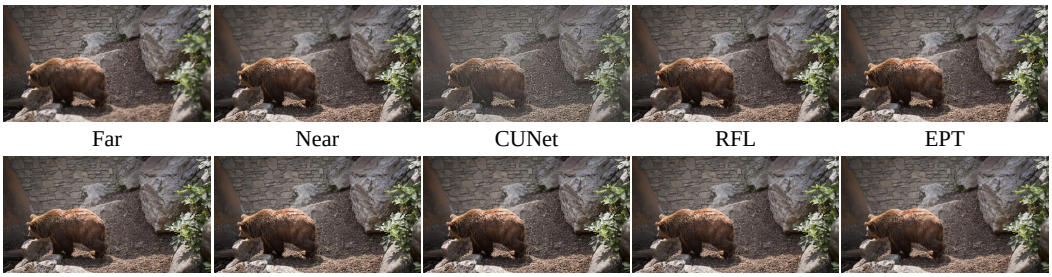

Figure 10: Visual comparison of fused results for multi-focus video fusion.

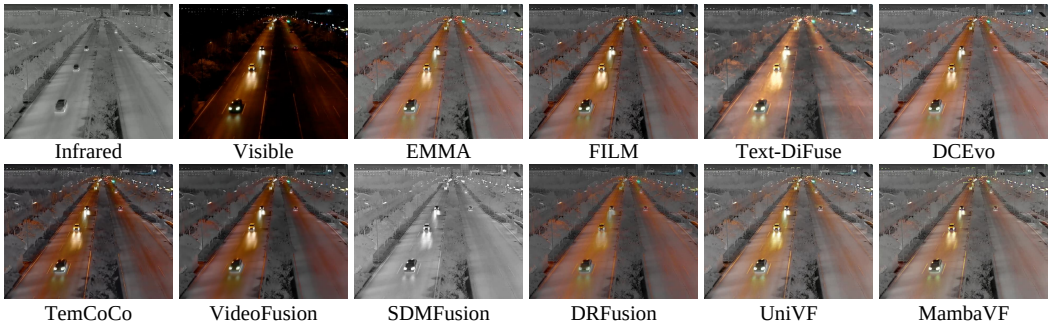

Figure 11: Visual comparison of fused results for infrared-visible video fusion.

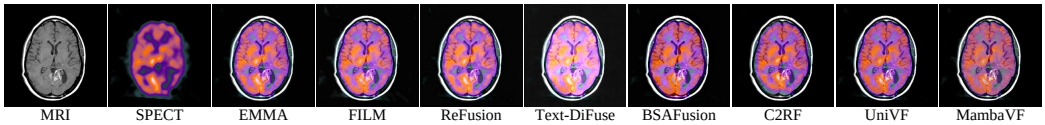

Figure 12: Visual comparison of fused results for medical video fusion.