# FREST: FREQUENCY REPRESENTATION ENHANCEMENT FOR SPATIOTEMPORAL ROBUSTNESS IN INFRARED AND VISIBLE VIDEO FUSION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Compared with single-frame image fusion, infrared and visible video fusion better reflects real-world sensing scenarios and supports continuous scene perception. However, existing spatial-domain methods struggle to separate degradation interference from informative structures and distinguish abnormal temporal fluctuations from valid motion, resulting in limited robustness to complex degradations and compromised temporal coherence. To address these issues, we propose **FREST**, which leverages frequency representations to suppress interference from degradation and abnormal temporal fluctuations while preserving informative content and valid motion. Specifically, it comprises two components: degradation-robust spatial-frequency modeling and frequency-aware spatiotemporal modeling. On the one hand, to suppress degradation-sensitive responses while preserving scene structures, we develop a degradation-robust spatial-frequency modeling framework to exploit the distinct spectral characteristics of degradation interference and scene structures. Specifically, we develop an adaptive frequency enhancement module and a spatial-frequency collaborative modeling strategy for dynamic modulation and effective cross-modal interaction. On the other hand, given that abnormal flickering and valid motion exhibit distinct response patterns in the joint spatiotemporal frequency domain, we construct a frequency domain temporal consistency module to model spectral changes across consecutive frames, thereby reducing unstable temporal fluctuations while preserving meaningful motion. Extensive experiments demonstrate that FREST produces more temporally coherent fused videos than state-of-the-art methods.

## 1 INTRODUCTION

Infrared and visible sensors provide complementary observations of the same scene. Infrared imaging captures thermal radiation and highlights targets under unfavorable illumination, whereas visible imaging records rich textures, edges, and structural details. Infrared and visible image fusion aims to integrate the complementary characteristics of these modalities, providing a more comprehensive visual representation than either modality alone and benefiting practical applications such as autonomous driving, surveillance and target perception (Zhao et al., 2023a; Zheng et al., 2024).

Considerable progress has been made in image fusion through long-range Transformer modeling, cross-modal attention, semantic guidance, and higher-order feature interactions (Ma et al., 2022; Li & Wu, 2024; Zhao et al., 2023a; Zheng et al., 2024). These techniques have substantially improved the extraction and integration of complementary spatial information, establishing a strong foundation for high-quality multimodal fusion. Compared with single-frame image fusion, infrared and visible video fusion further requires that the fused results remain reliable and coherent across continuous observations. Recent methods such as VideoFusion and TemCoCo have demonstrated that explicitly modeling temporal dependencies is essential for maintaining stable fusion results and supporting continuous scene understanding (Tang et al., 2026; Gong et al., 2025).

Nevertheless, existing video fusion methods still face two closely coupled challenges. First, infrared and visible videos are affected by heterogeneous degradations. For instance, infrared videos may contain sensor noise, whereas visible videos are susceptible to adverse weather and illumina-



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

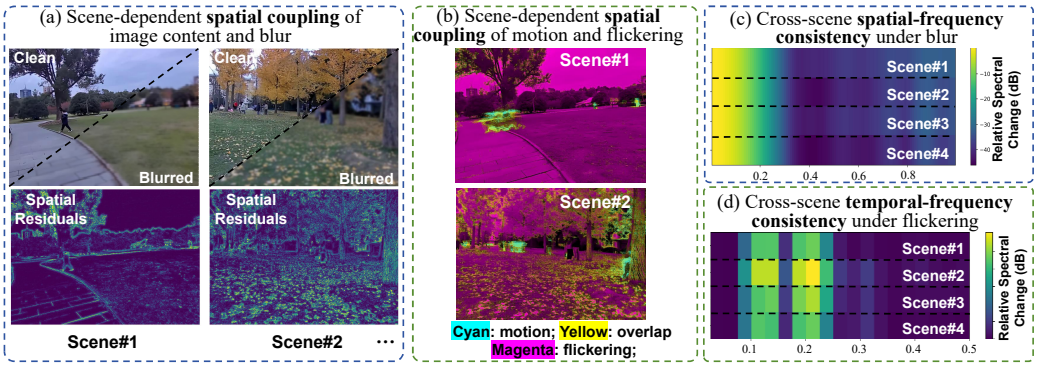



Figure 1: Illustration of scene-dependent spatial coupling and cross-scene frequency consistency. Four visually distinct scenes are selected for visualization. (a) Applying the same blur to different scenes produces spatial residuals that depend strongly on image content, whereas (c) the corresponding relative spatial-frequency changes remain highly consistent across scenes. (b) To simulate abnormal flickering, a global brightness fluctuation is added to each video sequence. Cyan, magenta, and yellow denote motion, flickering, and their overlap, respectively, showing that motion and flickering are highly entangled in the spatial domain. In contrast, (d) the relative temporal-frequency changes caused by flickering exhibit similar patterns across different motion sequences, indicating more consistent and discriminative frequency-domain responses.

tion variations. In the spatial domain, as shown in Figure 1(a), these degradations often produce responses that overlap with informative scene structures, making them difficult to identify and selectively suppress. Consequently, spatial feature processing may either retain degradation interference or inadvertently weaken salient thermal targets and fine visible textures. Second, undesirable inter-frame flickering may exhibit spatial variations similar to those caused by moving objects and normal scene changes, as shown in Figure 1(b). Spatial-domain temporal modeling may therefore fail to distinguish abnormal fluctuations from meaningful dynamics, leading to motion artifacts or over-smoothed details.

Frequency representations offer a complementary perspective on both challenges. For diverse degradations, spatial residuals are strongly coupled with scene content and vary across scenes. In contrast, relative spatial-frequency changes from the same degradation remain consistent, as shown in Figure 1(a) and (c). This consistency makes frequency representations a more content-insensitive basis for characterizing degradation. Meanwhile, Fourier transforms provide global receptive fields, enabling long-range dependency modeling without costly pairwise spatial-token interactions (Rao et al., 2021; Kong et al., 2023). For temporal consistency modeling, normal motion and abnormal flickering both appear as frame-wise intensity variations and strongly overlap in the spatial domain. However, flickering-induced relative temporal-frequency changes remain similar across diverse motion sequences, as shown in Figure 1(b) and (d), providing more discriminative cues for distinguishing abnormal flickering from valid motion. Therefore, 3D Fourier transforms are well suited to temporal feature modeling, enabling irregular fluctuations to be adaptively attenuated while preserving structured variations associated with genuine motion and scene changes.

Based on these observations, we propose **FREST**, a frequency-enhanced spatiotemporal representation learning framework for infrared and visible video fusion. The proposed framework establishes a unified frequency perspective to jointly address diverse degradation and temporal instability through two core designs: degradation-robust spatial-frequency modeling and frequency-aware spatiotemporal modeling. On the one hand, to construct degradation-robust modality representations, FREST introduces an adaptive frequency enhancement module that exploits the distinct spectral manifestations of degradation-induced disturbances and informative scene content. By separately refining spectral amplitude and phase and performing content-adaptive spectrum modulation, the module suppresses degradation-sensitive responses while preserving salient thermal targets and fine visible textures. A dual-domain collaborative strategy further integrates global spectral context with local spatial structures, while a lightweight frequency Transformer captures long-range dependencies through efficient element-wise frequency operations (Rao et al., 2021; Kong et al., 2023). On the

other hand, to maintain temporal coherence, FREST introduces a frequency-domain temporal consistency module that transforms consecutive video features into the joint spatiotemporal frequency domain and independently calibrates their real and imaginary components using lightweight 3D mappings. Rather than imposing rigid spatial smoothness, this module learns frequency-dependent spatiotemporal responses and attenuates irregular spectral fluctuations before cross-modal fusion. By combining degradation-robust spatial-frequency enhancement with frequency-aware spatiotemporal modeling, FREST generates fused videos with high-fidelity content, robustness to degradation, and coherent temporal dynamics.

The main contributions of this work are summarized as:

- We propose FREST, a unified spatiotemporal frequency framework for degradation-robust and temporally stable infrared–visible video fusion.
- We develop adaptive frequency enhancement and dual-domain collaborative modeling to exploit spatial-frequency and cross-modal complementarity.
- We introduce a frequency-domain temporal consistency module that models the real and imaginary components of the joint spatiotemporal spectrum, suppressing irregular flickering while preserving meaningful motion and scene changes.
- Extensive experiments demonstrate the effectiveness of FREST in fusion quality, degradation robustness, and temporal coherence.

## 2 Related Work

### 2.1 Infrared and Visible Image Fusion

Existing infrared and visible image fusion methods mainly include autoencoder, Transformer, and generative approaches. Autoencoder methods learn complementary features and reconstruct fused images through fusion strategies (Li & Wu, 2019; Xu et al., 2022). Transformer methods model long-range dependencies and cross-modal interactions (Ma et al., 2022; Zhao et al., 2023b). Generative methods improve visual quality and information preservation through adversarial learning and probabilistic modeling (Ma et al., 2019; Zhao et al., 2023c; Yi et al., 2024a; Zhang et al., 2026a). Recent methods further exploit degradation modeling, semantic guidance, and task, language, or cross-spectral priors (Tang et al., 2022; Liu et al., 2023; Zhao et al., 2023a; Yi et al., 2024b; Xu et al., 2026; Zhang et al., 2025; 2026b; Li et al., 2026a). However, these methods process image pairs and neglect temporal dependencies, thereby limiting their applicability to video fusion.

### 2.2 Infrared and Visible Video Fusion

Compared with static image fusion, infrared and visible video fusion must integrate complementary information within each frame while maintaining temporal coherence across consecutive frames. Existing methods employ temporal controllers, inter-frame interactions, optical flow alignment, feature propagation, and consistency constraints to reduce frame-wise fluctuations (Xie et al., 2024; Guo et al., 2024; Tang et al., 2026; Gong et al., 2025; Zhao et al., 2025b). Recent methods further introduce frequency-aware temporal perturbation or motion-selective interaction to improve spatiotemporal modeling (Li et al., 2026b;c). Nevertheless, jointly handling heterogeneous source degradation and temporal instability remains challenging. In the spatial domain, degradation responses may overlap with informative structures, while abnormal flickering can resemble valid motion. Our method addresses these issues within a unified frequency-oriented framework by combining degradation-robust spatial-frequency and frequency-aware spatiotemporal modeling.

## 3 Method

### 3.1 Overall Architecture

Given temporally synchronized infrared and visible video clips, our goal is to generate a fused video that preserves infrared thermal saliency, visible structural details, and coherent temporal dynamics.

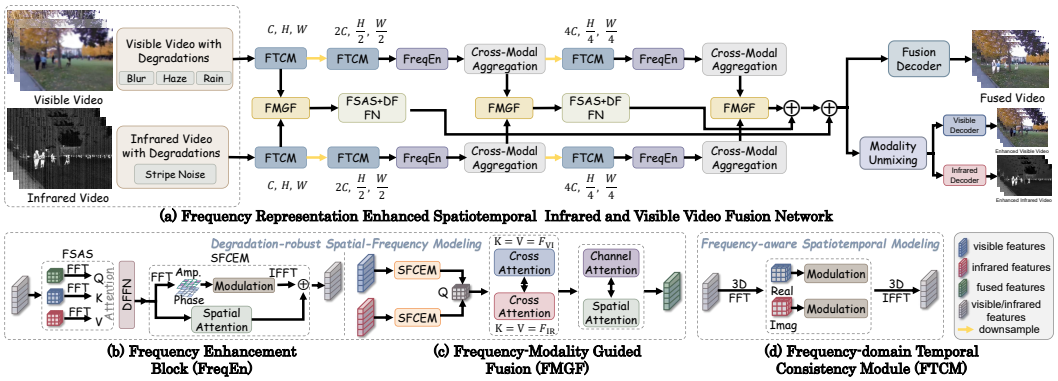



Figure 2: Architecture of FREST. (a) Overall architecture of the frequency-representation-enhanced network for spatiotemporal infrared and visible video fusion. (b) Frequency Enhancement Block, consisting of the frequency-domain-based self-attention solver, discriminative frequency-domain-based feed-forward network, and the proposed dual-domain collaborative enhancement module. (c) Frequency Modality-Guided Fusion. (d) Frequency-Domain Temporal Consistency Module.

The network also reconstructs the two source modalities to prevent degradation suppression from discarding source-supported information:

$$(\mathcal{V}^{\mathrm{f}}, \widehat{\mathcal{V}}^{\mathrm{ir}}, \widehat{\mathcal{V}}^{\mathrm{vi}}) = \mathcal{N}_\theta(\widetilde{\mathcal{V}}^{\mathrm{ir}}, \widetilde{\mathcal{V}}^{\mathrm{vi}}), \tag{1}$$

where $\widetilde{\mathcal{V}}^{x} = \{\widetilde{\mathbf{I}}_x^t\}_{t=1}^{T}$ is an input clip of modality $x \in \{\mathrm{ir}, \mathrm{vi}\}$, $T$ is the number of frames, $\mathcal{N}_\theta$ is the proposed network with parameters $\theta$, $\mathcal{V}^{\mathrm{f}}$ is the fused video, and $\widehat{\mathcal{V}}^{\mathrm{ir}}$ and $\widehat{\mathcal{V}}^{\mathrm{vi}}$ are the reconstructed infrared and visible videos, respectively.

As shown in Figure 2 (a), FREST adopts a three-scale encoder-decoder architecture and addresses robust video fusion through two complementary aspects: degradation-robust spatial-frequency modeling and frequency-aware spatiotemporal modeling. For degradation-robust spatial-frequency modeling, the infrared and visible branches first extract modality-specific features at multiple scales. At each scale, the *Frequency Enhancement Block* (FreqEn) adaptively enhances reliable modality representations by exploiting the distinct spectral manifestations of degradation interference and informative scene content. The enhanced infrared and visible features are subsequently exchanged and selectively integrated via cross-modal interaction and *Frequency Modality-Guided Fusion* (FMGF), thereby preserving frequency-enhanced cross-modal features. For frequency-aware spatiotemporal modeling, the *Frequency-Domain Temporal Consistency Module* (FTCM) models video features in the joint spatiotemporal frequency domain to attenuate irregular fluctuations while preserving meaningful scene dynamics. The decoder progressively reconstructs informative, temporally coherent fused videos through coarse-to-fine aggregation and local cross-frame interactions.

After degradation-robust spatial-frequency enhancement and frequency-aware spatiotemporal modeling, the enhanced features are fed into the decoder. The decoder follows a coarse-to-fine reconstruction process (Tang et al., 2026). Deep fused features are progressively upsampled and combined with the corresponding finer-scale features, while stacked bidirectional co-attention blocks aggregate adjacent-frame information. At the finest scale, a video decomposition unit separates the decoded representation into infrared-related and visible-related features. Two auxiliary decoders reconstruct the enhanced modalities, and their refined aggregation is decoded into the final fused video. The joint fusion and reconstruction design retains enhanced thermal targets and visible textures.

### 3.2 Degradation-Robust Spatial-Frequency Modeling

In the spatial domain, degradation responses and informative scene structures remain entangled within each frame. To obtain degradation-robust representations, we introduce a FreqEn and an FMGF to enhance frequency representations and facilitate frequency-guided cross-modal fusion, as shown in Figure 2 (b) and (c).

**FreqEn.** This design consists of the *frequency-domain-based self-attention solver* (FSAS), the *discriminative frequency-domain-based feed-forward network* (DFFN), and the *Spatial-Frequency Collaborative Enhancement Module* (SFCEM). FSAS and DFFN follow the efficient frequency-domain formulation of (Kong et al., 2023), whereas SFCEM is developed to facilitate the mutual enhancement of both the spatial and frequency domains.

Given a modality-specific feature $\mathbf{X}$, FSAS first projects query, key, and value features $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$. Let $\widehat{\mathbf{Q}} = \mathcal{F}_{2\mathrm{D}}(\mathcal{P}(\mathbf{Q}))$ and $\widehat{\mathbf{K}} = \mathcal{F}_{2\mathrm{D}}(\mathcal{P}(\mathbf{K}))$. Following FFTformer (Kong et al., 2023), their patch-wise correlation is computed in the frequency domain as

$$\mathbf{C}_{qk} = \mathcal{P}^{-1}\Big\{\mathcal{F}_{2\mathrm{D}}^{-1}\left(\widehat{\mathbf{Q}} \odot \overline{\widehat{\mathbf{K}}}\right)\Big\}, \tag{2}$$

where $\mathcal{P}$ and $\mathcal{P}^{-1}$ denote patch partition and recovery, $\mathcal{F}_{2\mathrm{D}}$ and $\mathcal{F}_{2\mathrm{D}}^{-1}$ are the 2D FFT and its inverse, $\odot$ denotes element-wise multiplication, and the overline denotes complex conjugation. The correlation response is normalized before modulating the value feature, giving

$$\mathbf{Y} = \mathbf{X} + \mathcal{H}_o(\mathbf{V} \odot \mathcal{L}(\mathbf{C}_{qk})), \tag{3}$$

where $\mathcal{L}$ denotes normalization and $\mathcal{H}_o$ is an output projection. This formulation approximates attention-like long-range interaction without explicitly constructing a quadratic spatial attention matrix.

Subsequently, DFFN applies a learnable spectral filter to select task-relevant frequency components. The process can be represented as:

$$\mathbf{U} = \mathcal{P}^{-1}\big\{\mathcal{F}_{2\mathrm{D}}^{-1}(\mathbf{\Theta} \odot \mathcal{F}_{2\mathrm{D}}\big(\mathcal{P}(\mathcal{H}_p(\mathbf{Y}))\big))\big\}, \tag{4}$$

where $\mathbf{Y}$ is the input to DFFN, $\mathcal{H}_p$ is an input projection, and $\mathbf{\Theta}$ is a learnable spectral filter. Finally, DFFN is defined as:

$$\mathrm{DFFN}(\mathbf{Y}) = \mathbf{Y} + \mathcal{H}_d(\phi(\mathbf{U}_1) \odot \mathbf{U}_2), \tag{5}$$

where $[\mathbf{U}_1, \mathbf{U}_2]$ is obtained by channel-wise splitting after depth-wise transformation, $\phi$ is a nonlinear activation, and $\mathcal{H}_d$ is an output projection.

Frequency filtering provides global spectral regulation but may weaken spatially localized structures. We therefore design SFCEM to combine adaptive amplitude–phase modulation with local spatial processing. For an input feature $\mathbf{Y}$, let $\mathcal{F}_{2\mathrm{D}}(\mathbf{Y}) = \mathbf{A} \odot e^{\mathrm{i}\mathbf{P}}$, where $\mathbf{A}$ and $\mathbf{P}$ are the spectral amplitude and phase. We modulate them by $\widetilde{\mathbf{A}} = \mathbf{g}_A \odot [\mathbf{A} + \mathcal{H}_A(\mathbf{A})]$ and $\widetilde{\mathbf{P}} = \mathbf{P} + \mathcal{H}_P(\mathbf{P})$, where $\mathcal{H}_A$ and $\mathcal{H}_P$ refine the two spectral components and $\mathbf{g}_A$ is an amplitude gate. The reconstructed frequency-enhanced feature is:

$$\mathbf{Y}_{re} = \mathrm{Re}\Big\{\mathcal{F}_{2\mathrm{D}}^{-1}(\widetilde{\mathbf{A}} \odot e^{\mathrm{i}\widetilde{\mathbf{P}}})\Big\}, \tag{6}$$

where $\mathbf{Y}_{re}$ is the reconstructed frequency-enhanced feature.

The SFCEM is defined as:

$$\mathrm{SFCEM}(\mathbf{Y}) = \mathbf{Y} + \mathcal{H}_c(\mathrm{Cat}[\mathbf{Y}_s, \mathbf{g}_f \odot \mathbf{Y}_{re}]), \tag{7}$$

where $\mathbf{Y}_s = \mathcal{H}_s(\mathbf{Y})$ is the local spatial feature, $\mathbf{g}_f$ is a frequency gate, $\mathrm{Cat}[\cdot]$ denotes channel concatenation, and $\mathcal{H}_c$ projects the concatenated dual-domain feature. The frequency branch regulates degradation-sensitive spectral responses, while the spatial branch preserves textures and localized thermal structures.

**FMGF.** Independent enhancement cannot fully exploit the complementary information between infrared and visible modalities. Moreover, direct addition or concatenation may propagate modality-specific degradation into the fused representation. As illustrated in Figure 2 (a), we first perform gated cross-modality difference enhancement followed by feature aggregation. For modality $x$ and another modality $\bar{x}$, the gated complementary difference is defined as:

$$\mathbf{G}_x = \sigma(\mathcal{H}_g(\mathrm{Cat}[\mathbf{E}_x, \mathbf{E}_{\bar{x}}])), \tag{8}$$

$$\mathbf{D}_x = \mathbf{G}_x \odot (\mathbf{E}_{\bar{x}} - \mathbf{E}_x), \tag{9}$$

where $\sigma$ is the sigmoid function, $\mathcal{H}_g$ predicts the gate $\mathbf{G}_x$, and $\mathbf{D}_x$ retains complementary responses supported by the other modality.

The interacted feature is obtained by:

$$\widetilde{\mathbf{F}}_x = \mathcal{C}(\mathcal{A}(\mathbf{E}_x, \mathbf{D}_x), \mathbf{E}_x), \tag{10}$$

where $\mathcal{A}$ denotes cross-attention and $\mathcal{C}$ denotes contribution-aware aggregation with channel and spatial attention. This interaction introduces complementary evidence while reducing the transfer of unreliable modality-specific differences.

The interacted features are then fused by FMGF, shown in Figure 2 (c). Each modality is first refined as $\mathbf{S}_x = \mathrm{SFCEM}(\widetilde{\mathbf{F}}_x)$, and the common query is $\mathbf{Q}_c = \mathcal{H}_q(\mathrm{Cat}[\mathbf{S}_{\mathrm{ir}}, \mathbf{S}_{\mathrm{vi}}])$. The two guided responses are:

$$\mathbf{A}_x = \mathcal{A}(\mathbf{Q}_c, \mathbf{S}_x), \qquad x \in \{\mathrm{ir}, \mathrm{vi}\}, \tag{11}$$

where $\mathcal{H}_q$ constructs the common query $\mathbf{Q}_c$ and $\mathcal{A}$ is the shared cross-attention mapping.

The output of FMGF is defined as:

$$\mathbf{F}_f = \mathcal{H}_f[\mathcal{H}_{\mathrm{ch}}(\mathbf{A}_{\mathrm{ir}}) + \mathcal{H}_{\mathrm{sp}}(\mathbf{A}_{\mathrm{vi}})], \tag{12}$$

where $\mathcal{H}_{\mathrm{ch}}$ and $\mathcal{H}_{\mathrm{sp}}$ denote channel and spatial attention refinement, respectively, and $\mathcal{H}_f$ is the output projection. We denote the complete mapping by $\mathbf{F}_f = \mathrm{FMGF}(\widetilde{\mathbf{F}}_{\mathrm{ir}}, \widetilde{\mathbf{F}}_{\mathrm{vi}})$.

By using a shared query, FMGF selects complementary infrared and visible evidence under a common reference. The subsequent channel and spatial refinement preserves salient thermal targets and visible structures while limiting degradation-dominated responses.

### 3.3 Frequency-Aware Spatiotemporal Modeling

Degradation-induced flickering closely resembles valid object motion in local temporal dynamics. Consequently, spatial processing struggles to distinguish abnormal fluctuations from genuine scene motion. To address this limitation, we introduce an FTCM to model features in the joint spatiotemporal frequency domain, as illustrated in Figure 2 (d).

Given a modality feature $\mathbf{F}_x^l \in \mathbb{R}^{C_l \times T \times H_l \times W_l}$ at scale $l$, its spatiotemporal spectrum is computed using a 3D Fast Fourier Transform (3D FFT) along the dimensions $(T, H_l, W_l)$:

$$\mathbf{Z}_x^l = \mathcal{F}_{\mathrm{3D}}(\mathbf{F}_x^l) = \mathbf{R}_x^l + \mathrm{i}\mathbf{I}_x^l, \tag{13}$$

where $\mathcal{F}_{\mathrm{3D}}$ represents the 3D FFT process, $\mathbf{Z}_x^l$ is the complex-valued spectrum, $\mathbf{R}_x^l$ and $\mathbf{I}_x^l$ are real and imaginary components, and $\mathrm{i}$ is the imaginary unit.

To enable adaptive refinement of spectral responses associated with inter-frame variations, the real and imaginary components are independently calibrated as:

$$\widetilde{\mathbf{Z}}_x^l = \mathcal{T}_R^l(\mathbf{R}_x^l) + \mathrm{i}\mathcal{T}_I^l(\mathbf{I}_x^l), \tag{14}$$

where $\mathcal{T}_R^l$ and $\mathcal{T}_I^l$ denote independent three-dimensional convolutional transformations applied to the real and imaginary components, respectively, and $\widetilde{\mathbf{Z}}_x^l$ represents the calibrated complex-valued spectrum. By aggregating neighboring spectral coefficients along the temporal and spatial frequency axes, these transformations capture local dependencies among adjacent frequency components and provide learnable modulation of their responses.

Finally, FTCM transforms the output to the spatial domain:

$$\overline{\mathbf{F}}_x^l = \mathrm{Re}\Big\{\mathcal{F}_{\mathrm{3D}}^{-1}(\widetilde{\mathbf{Z}}_x^l)\Big\}, \tag{15}$$

where $\mathcal{F}_{\mathrm{3D}}^{-1}$ is the inverse 3D FFT and $\mathrm{Re}\{\cdot\}$ extracts the real component.

### 3.4 Loss Functions

We adopt the intensity, gradient, color, scene-fidelity, and variation-consistency losses from VideoFusion (Tang et al., 2026). The overall objective is

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{int}}\mathcal{L}_{\mathrm{int}} + \lambda_{\mathrm{grad}}\mathcal{L}_{\mathrm{grad}} + \lambda_{\mathrm{color}}\mathcal{L}_{\mathrm{color}} + \lambda_{\mathrm{sf}}\mathcal{L}_{\mathrm{sf}} + \lambda_{\mathrm{var}}\mathcal{L}_{\mathrm{var}}, \tag{16}$$

where $\lambda$ denotes the corresponding loss weights.

# 4 EXPERIMENTS

We evaluate FREST from four complementary aspects. We first describe the experimental data, evaluation metrics, implementation details, and competing methods. We then assess fusion quality and degradation robustness through quantitative evaluation and qualitative comparisons under both synthetic and real-world degradations. Finally, we analyze temporal consistency and conduct ablation studies to validate the contribution of each component.

## 4.1 DATASETS AND IMPLEMENTATION

**Datasets.** We used two infrared–visible video datasets for the experiments, M3SVD (Tang et al., 2026) and HDO (Xie et al., 2024). M3SVD contains 220 video sequences, each comprising 100 to 1000 frames, while HDO includes 26 video sequences, each with approximately 150 to 200 frames.

**Evaluation Metrics.** We employ four metrics to evaluate spatial fusion quality and temporal consistency. MI (Qu et al., 2002) measures information retained from the source modalities, while SSIM (Wang et al., 2004) evaluates structural similarity. For temporal quality, flowD (Gong et al., 2025) measures inter-frame temporal discrepancy, and Bi-Directional Self-Warping Error (BiSWE) (Zhao et al., 2025b) evaluates temporal consistency through bidirectional self-warping. Higher MI and SSIM are better, whereas lower flowD and BiSWE indicate better temporal consistency.

**Implementation Details.** Training samples are randomly cropped to $128 \times 128$ and augmented by horizontal and vertical flips and random rotations. FREST is trained on M3SVD for 10 epochs using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay $1 \times 10^{-5}$. The learning rate is initially set to $1 \times 10^{-4}$ and cosine annealed to $1 \times 10^{-6}$. The loss weights are set to $\lambda_{\text{int}} = 15$, $\lambda_{\text{grad}} = 1$, $\lambda_{\text{color}} = 100$, $\lambda_{\text{sf}} = 1$, and $\lambda_{\text{var}} = 1$. The three stages use $[2, 2, 4]$ blocks with $[32, 64, 128]$ channels. All experiments are conducted on NVIDIA GeForce RTX 3090 GPUs and 2.40 GHz Intel Xeon Silver 4210R CPUs.

For all-in-one training, degraded infrared videos are uniformly paired with blurred, hazy, rainy, or clean visible videos. Blur uses a $12 \times 12$ isotropic Gaussian kernel with $\sigma \in [0.5, 0.8]$, fixed within each sequence. Haze is synthesized using the atmospheric scattering model with a temporal smoothing factor of 0.93. Rain uses density $[0.08, 0.09]$, streak length $[40, 45]$, and intensity $[0.9, 1.1]$. The current and shifted previous rain maps are blended with weights of 0.7 and 0.3. Gaussian noise and column-stripe noise are added to infrared videos, and a bias field with intensity 0.2 is applied with a probability of 0.5.

**Compared Methods.** We compare FREST with four image fusion methods, including LRRNet (Li et al., 2023), CrossFuse (Li & Wu, 2024), GIFNet (Cheng et al., 2025), and FreeFusion (Zhao et al., 2025a), as well as three video fusion methods: TemCoCo (Gong et al., 2025), UniVF (Zhao et al., 2025b), and MAVFusion (Li et al., 2026c). For a fair comparison under degraded inputs, we also employ representative video restoration algorithms to pre-enhance the source videos before fusion, including TURTLE (Ghasemabadi et al., 2024) for deblurring, dehazing, and deraining visible videos, and MDIVDNet (Cai et al., 2024) for infrared denoising.

## 4.2 FUSION PERFORMANCE COMPARISON

**Qualitative Comparison.** Figure 3 compares fusion results under clean and compound degradation settings. Under clean inputs in Figure 3(a), several image-based baselines lose local structures, while the video baselines TemCoCo, UniVF, and MAVFusion tend to smooth fine textures to different degrees. Under infrared noise with visible blur in (b), competing methods retain residual blur or noise and may introduce local color distortion, whereas FREST preserves sharper contours and cleaner textures. In the haze case of (c), FREST preserves clearer moving targets, finer branch structures, and a more natural appearance. In the low-light rain case of (d), competing methods exhibit degraded window textures or weakened building edges, while FREST maintains clearer structures and salient thermal targets. Overall, FREST is more robust to compound degradations while preserving complementary information from both modalities.

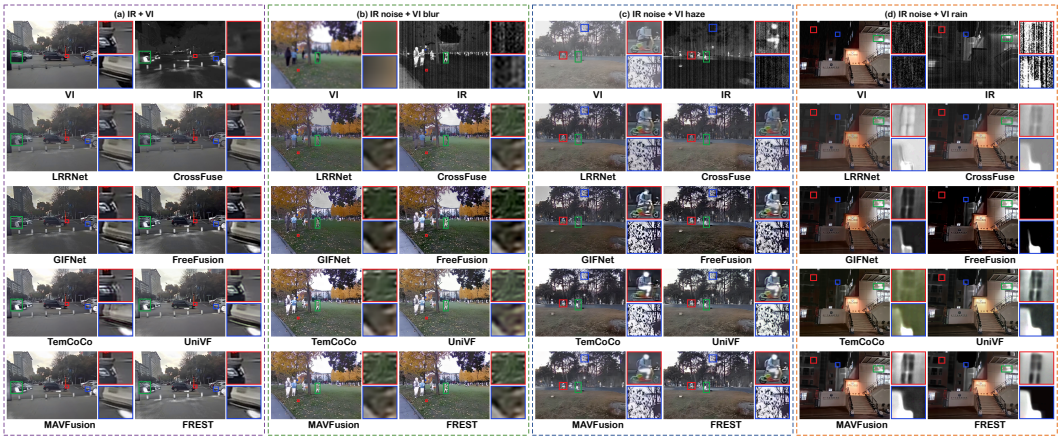



Figure 3: Qualitative comparison under (a) clean inputs, (b) infrared noise and visible blur, (c) infrared noise and visible haze, and (d) infrared noise and visible rain. Zoomed regions in (d) are brightened by 5× for visibility.

**Real World Degradation.** Figure 4 further evaluates the methods on real world degraded inputs, where degradation patterns are more complex and unpredictable than the synthetic corruptions used during training. Compared with competing image and video fusion methods, FREST more effectively suppresses visible degradation artifacts while preserving salient structures and informative details, thereby demonstrating stronger and more consistent robustness under the tested real world conditions.

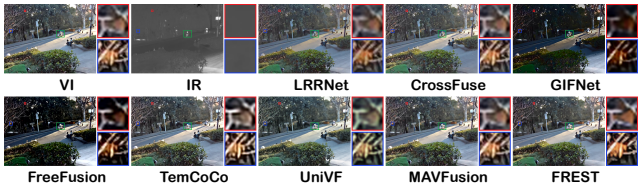



Figure 4: Qualitative video fusion results with real-world degradation.

**Quantitative Comparison.** Table 2 reports the results on 17 M3SVD test sequences and the aligned HDO test set under four evaluation settings. Across the eight dataset–setting combinations, FREST obtains the best MI and SSIM in seven cases, the best flowD in six cases, and the best BiSWE in seven cases. On M3SVD, FREST ranks first on all four metrics for the IR En.+VI En. and blur settings. Under haze, it achieves the best MI and SSIM and the second-best flowD and BiSWE; under rain, it obtains the best SSIM, flowD, and BiSWE and the second-best MI. On HDO, FREST achieves the best results on all four metrics under blur, haze, and rain. For IR En.+VI En., it achieves the best MI and BiSWE and the second-best SSIM and flowD. These results show that FREST improves spatial fusion quality while maintaining strong temporal stability across datasets and degradation types.

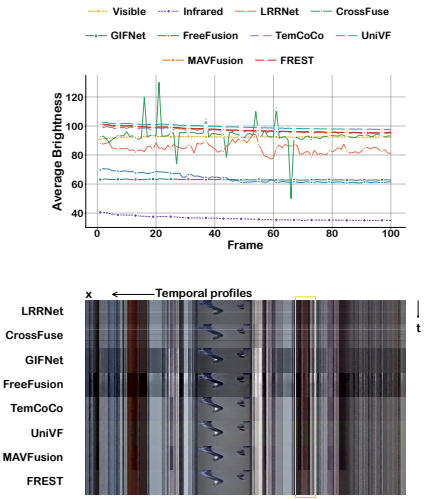



Figure 5: Temporal consistency comparison.

**Temporal Consistency.** We evaluate temporal consistency using flowD and BiSWE in Table 2, complemented by temporal profiles in Figure 5, constructed by stacking the same horizontal scanline from 100 consecutive frames. FREST achieves the lowest flowD in six of eight dataset–setting combinations and the lowest BiSWE in seven, ranking second in the remaining cases. Under blur and rain, it consistently outperforms TemCoCo, UniVF, and MAVFusion on both datasets. The temporal profiles further show smooth structural trajectories and stable inter-frame intensity transitions while preserving local detail, corroborating the quantitative results.

Table 2: Quantitative comparison on M3SVD and HDO under four evaluation settings. The best and second-best results are highlighted in bold and underlined, respectively, within each setting and dataset.

| | | | | M3SVD | | | | HDO | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Scenario** | **Type** | **Method** | **Pub.** | **MI ↑** | **SSIM ↑** | **flowD ↓** | **BiSWE ↓** | **MI ↑** | **SSIM ↑** | **flowD ↓** | **BiSWE ↓** |
| IR En. + VI En. | Image | LRRNet | *TPAMI'23* | 2.467 | 0.6585 | 5.794 | 4.438 | 1.789 | 0.6220 | **6.361** | 6.069 |
| | | CrossFuse | *InfFus'24* | 2.801 | 0.6504 | 10.496 | 9.735 | 2.096 | 0.6640 | 9.499 | 9.012 |
| | | GIFNet | *CVPR'25* | 1.704 | 0.6655 | 5.436 | 3.391 | 1.562 | 0.6582 | 8.417 | 8.175 |
| | | FreeFusion | *TPAMI'25* | 1.844 | 0.5783 | 7.669 | 5.286 | 1.726 | 0.6356 | 9.394 | 9.231 |
| | Video | TemCoCo | *ICCV'25* | 2.845 | 0.6353 | 7.799 | 5.076 | 1.953 | 0.6276 | 8.332 | 6.961 |
| | | UniVF | *NeurIPS'25* | 3.362 | 0.6620 | 5.374 | 2.770 | 2.326 | **0.6776** | 7.819 | 7.178 |
| | | MAVFusion | *ECCV'26* | 3.371 | 0.6572 | 5.780 | 3.297 | 2.142 | 0.6648 | 8.168 | 7.555 |
| | | **FREST** | – | **3.617** | **0.6664** | **5.012** | **2.597** | **2.469** | 0.6691 | 6.606 | **5.741** |
| IR Noise + VI Blur | Image | LRRNet | *TPAMI'23* | 2.623 | 0.6750 | 6.287 | 5.338 | 1.661 | 0.5915 | 7.636 | 8.317 |
| | | CrossFuse | *InfFus'24* | 2.881 | 0.6689 | 7.565 | 7.018 | 1.705 | 0.5856 | 12.017 | 13.563 |
| | | GIFNet | *CVPR'25* | 1.799 | 0.6791 | 5.704 | 4.158 | 1.506 | 0.5585 | 9.399 | 10.691 |
| | | FreeFusion | *TPAMI'25* | 1.955 | 0.6275 | 7.085 | 5.259 | 1.608 | 0.5323 | 10.662 | 12.401 |
| | Video | TemCoCo | *ICCV'25* | 3.163 | 0.6641 | 7.873 | 5.334 | 1.795 | 0.5860 | 7.531 | 6.726 |
| | | UniVF | *NeurIPS'25* | 3.398 | 0.6742 | 4.813 | 2.519 | 1.872 | 0.5948 | 8.274 | 9.043 |
| | | MAVFusion | *ECCV'26* | 3.267 | 0.6695 | 5.255 | 3.175 | 1.716 | 0.5755 | 8.966 | 9.901 |
| | | **FREST** | – | **3.585** | **0.6792** | **4.583** | **2.486** | **2.064** | **0.6112** | **6.188** | **6.148** |
| IR Noise + VI Haze | Image | LRRNet | *TPAMI'23* | 2.527 | 0.6740 | 6.489 | 5.546 | 1.694 | 0.6069 | 7.094 | 7.337 |
| | | CrossFuse | *InfFus'24* | 2.645 | 0.6662 | 8.606 | 8.269 | 1.792 | 0.6130 | 21.508 | 25.026 |
| | | GIFNet | *CVPR'25* | 1.771 | 0.6793 | 5.907 | 4.378 | 1.523 | 0.5849 | 8.728 | 9.353 |
| | | FreeFusion | *TPAMI'25* | 1.930 | 0.6315 | 7.348 | 5.584 | 1.670 | 0.5721 | 9.768 | 10.632 |
| | Video | TemCoCo | *ICCV'25* | 2.833 | 0.6517 | 7.811 | 5.392 | 1.710 | 0.5729 | 7.613 | 6.938 |
| | | UniVF | *NeurIPS'25* | 3.101 | 0.6715 | **4.934** | **2.734** | 1.923 | 0.6202 | 7.800 | 8.075 |
| | | MAVFusion | *ECCV'26* | 3.018 | 0.6669 | 5.554 | 3.557 | 1.792 | 0.6029 | 8.511 | 8.779 |
| | | **FREST** | – | **3.188** | **0.6803** | 4.957 | 2.986 | **2.170** | **0.6257** | **6.056** | **5.717** |
| IR Noise + VI Rain | Image | LRRNet | *TPAMI'23* | 2.509 | 0.6713 | 6.940 | 6.228 | 1.737 | 0.5989 | 7.081 | 7.376 |
| | | CrossFuse | *InfFus'24* | 2.639 | 0.6638 | 8.667 | 8.557 | 1.863 | 0.6193 | 13.035 | 14.518 |
| | | GIFNet | *CVPR'25* | 1.785 | 0.6774 | 6.337 | 5.129 | 1.568 | 0.5903 | 8.369 | 8.892 |
| | | FreeFusion | *TPAMI'25* | 1.931 | 0.6323 | 7.934 | 6.587 | 1.704 | 0.5841 | 9.431 | 10.020 |
| | Video | TemCoCo | *ICCV'25* | 2.908 | 0.6551 | 7.918 | 5.504 | 1.781 | 0.5871 | 7.587 | 6.860 |
| | | UniVF | *NeurIPS'25* | **3.017** | 0.6708 | 5.521 | 3.580 | 1.970 | 0.6224 | 7.737 | 7.913 |
| | | MAVFusion | *ECCV'26* | 2.863 | 0.6660 | 6.217 | 4.597 | 1.831 | 0.6067 | 8.299 | 8.447 |
| | | **FREST** | – | 2.959 | **0.6795** | **4.960** | **3.155** | **2.178** | **0.6238** | **5.903** | **5.594** |

### 4.3 Ablation Studies

Table 1: Ablation results on M3SVD.

| | **MI ↑** | **SSIM ↑** | **flowD ↓** | **BiSWE ↓** |
|---|---|---|---|---|
| w/o FreqEn | 2.993 | 0.629 | 5.417 | 3.456 |
| w/o SFCEM | 2.904 | 0.631 | 5.298 | 3.349 |
| w/o FMGF | 2.878 | 0.630 | 5.362 | 3.585 |
| w/o FTCM | 3.036 | 0.634 | 7.196 | 3.880 |
| w/o $\mathcal{L}_{\text{grad}}$ | 3.037 | 0.633 | 5.996 | 3.248 |
| w/o $\mathcal{L}_{\text{color}}$ | 2.992 | 0.628 | 9.893 | 3.646 |
| **FREST** | **3.083** | **0.635** | **5.163** | **3.155** |

Table 1 verifies the contributions of the proposed components and training objectives. Among the architectural variants, removing FMGF leads to the most pronounced degradation in spatial fusion quality, highlighting the importance of cross-modal feature interaction, while removing FTCM notably worsens temporal consistency. FreqEn and SFCEM also contribute to a better spatial–temporal balance. For the loss terms, both $\mathcal{L}_{\text{grad}}$ and $\mathcal{L}_{\text{color}}$ contribute to fusion quality and temporal stability. Overall, the full FREST consistently achieves the best performance across spatial quality and temporal consistency.

## 5 Conclusion

This work has presented **FREST**, a frequency-enhanced spatiotemporal framework for robust infrared and visible video fusion. On the one hand, a Frequency Enhancement Block with spatial-frequency collaborative enhancement regulates degradation-sensitive spectral responses while preserving textures and thermal structures. At the same time, Frequency Modality-Guided Fusion integrates complementary cross-modal information. On the other hand, a Frequency-Domain Temporal Consistency Module models consecutive features in the joint spatiotemporal frequency domain, suppressing irregular fluctuations while retaining meaningful motion. Experiments on M3SVD and HDO demonstrate the advantages of FREST over image-based and video-based fusion methods in handling compound degradations and producing informative, clear, and temporally coherent results.

## AI USE STATEMENT

Generative AI tools were used to assist with language polishing, LaTeX formatting checks, and bibliography-metadata verification. All AI-assisted suggestions were reviewed by the authors against the manuscript, implementation, experimental records, and primary publication sources as applicable. The authors take responsibility for the final content of the paper, including all technical claims, equations, citations, experimental results, and presentation.

## ETHICS STATEMENT

This work studies infrared–visible video fusion using existing research benchmarks and does not introduce new human-subject experiments or collect new personally identifiable data. Infrared–visible sensing and fusion can support applications such as autonomous systems and surveillance; downstream deployment should therefore consider privacy, data-governance, and misuse risks appropriate to the application context. Users of the benchmark data and released artifacts should follow the corresponding dataset and software licenses.

## REPRODUCIBILITY STATEMENT

The model architecture and frequency-domain operations are described in Section 3, including the training objective in Section 3.4. Section 4.1 reports the datasets, degradation synthesis settings, optimization hyperparameters, model configuration, and hardware used in our experiments. Section 4.3 and Appendix A provide component-level ablations, and Appendix B reports computational efficiency. These details, together with the cited public benchmarks and baseline implementations, are intended to support independent reproduction of the reported results.

## REFERENCES

Lijing Cai, Xiangyu Dong, Kailai Zhou, and Xun Cao. Exploring Video Denoising in Thermal Infrared Imaging: Physics-Inspired Noise Generator, Dataset, and Model. *IEEE Transactions on Image Processing*, 33:3839–3854, 2024.

Chunyang Cheng, Tianyang Xu, Zhenhua Feng, Xiaojun Wu, Zhangyong Tang, Hui Li, Zeyang Zhang, Sara Atito, Muhammad Awais, and Josef Kittler. One Model for ALL: Low-Level Task Interaction Is a Key to Task-Agnostic Image Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28102–28112, 2025.

Amirhosein Ghasemabadi, Muhammad Kamran Janjua, Mohammad Salameh, and Di Niu. Learning Truncated Causal History Model for Video Restoration. In *Advances in Neural Information Processing Systems*, volume 37, pp. 27584–27615, 2024.

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.

Xiaoming Guo, Fengbao Yang, and Linna Ji. DFTP: A Mimic Fusion Method Driven by Difference Feature Time-Phase Perception for Infrared and Visible Video. *IEEE Transactions on Instrumentation and Measurement*, 73:1–20, 2024.

Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, and Jinshan Pan. Efficient Frequency Domain-Based Transformers for High-Quality Image Deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5886–5895, 2023.

Hui Li and Xiao-Jun Wu. DenseFuse: A Fusion Approach to Infrared and Visible Images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019.

Hui Li and Xiao-Jun Wu. CrossFuse: A Novel Cross Attention Mechanism Based Infrared and Visible Image Fusion Approach. *Information Fusion*, 103:102147, 2024.

Hui Li, Tianyang Xu, Xiao-Jun Wu, Jiwen Lu, and Josef Kittler. LRRNet: A Novel Representation Learning Guided Fusion Network for Infrared and Visible Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11040–11052, 2023.

Jing Li, Yifan Wang, Jiafeng Yan, Renlong Zhang, and Bin Yang. MdaIF: Robust One-Stop Multi-Degradation-Aware Image Fusion with Language-Driven Semantics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6226–6234, 2026a.

Xilai Li, Chusheng Fang, and Xiaosong Li. FTPFusion: Frequency-Aware Infrared and Visible Video Fusion with Temporal Perturbation. *arXiv preprint arXiv:2604.01900*, 2026b.

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. In *Proceedings of the European Conference on Computer Vision*, 2026c.

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-Interactive Feature Learning and a Full-Time Multi-Modality Benchmark for Image Fusion and Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8115–8124, 2023.

Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. FusionGAN: A Generative Adversarial Network for Infrared and Visible Image Fusion. *Information Fusion*, 48:11–26, 2019.

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.

Guohui Qu, Dali Zhang, and Pingfan Yan. Information Measure for Performance of Image Fusion. *Electronics Letters*, 38(7):313–315, 2002.

Yongming Rao, Wenliang Zhao, Zheng Zhu, Jiwen Lu, and Jie Zhou. Global Filter Networks for Image Classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 980–993, 2021.

Linfeng Tang, Jiteng Yuan, and Jiayi Ma. Image Fusion in the Loop of High-Level Vision Tasks: A Semantic-Aware Real-Time Infrared and Visible Image Fusion Network. *Information Fusion*, 82: 28–42, 2022.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19559–19569, 2026.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

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, 2022.

Xingxin Xu, Bing Cao, Dongdong Li, Qinghua Hu, and Pengfei Zhu. Dream-if: Dynamic relative enhancement for image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11442–11450, 2026.

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, 2024a.

Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-IF: Leveraging Semantic Text Guidance for Degradation-Aware and Interactive Image Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27026–27035, 2024b.

Hao Zhang, Lei Cao, Xuhui Zuo, Zhenfeng Shao, and Jiayi Ma. OmniFuse: Composite Degradation-Robust Image Fusion with Language-Driven Semantics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7577–7595, 2025.

Hao Zhang, Meiqi Gong, Douyu Wu, and Jiayi Ma. DPAFuse: A Dual-Space Probabilistic Adversarial Image Fusion Framework with Robust Coding Embedding. *Fundamental Research*, 6(4): 2587–2601, 2026a.

Hao Zhang, Yanping Zha, Zizhuo Li, Meiqi Gong, and Jiayi Ma. MagicFuse: Single Image Fusion for Visual and Semantic Reinforcement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26551–26560, 2026b.

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.

Wenda Zhao, Hengshuai Cui, Haipeng Wang, You He, and Huchuan Lu. FreeFusion: Infrared and Visible Image Fusion via Cross Reconstruction Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):8040–8056, 2025a.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8082–8093, 2023c.

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 *Advances in Neural Information Processing Systems*, volume 38, pp. 142598–142629, 2025b.

Naishan Zheng, Man Zhou, Jie Huang, Junming Hou, Haoying Li, Yuan Xu, and Feng Zhao. Probing Synergistic High-Order Interaction in Infrared and Visible Image Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26384–26395, 2024.

## A Qualitative Ablation Results

We further provide qualitative ablation results to visually examine the contribution of each component in FREST. Figure 6 compares the complete model with variants obtained by removing individual architectural components or training objectives.

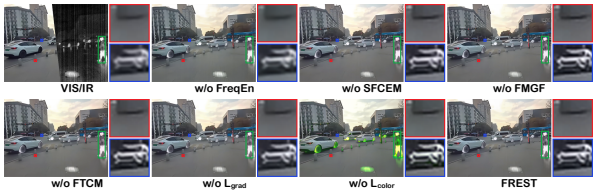



Figure 6: Qualitative ablation comparison on M3SVD. Enlarged regions are provided to facilitate comparison of degradation suppression, target preservation, and local structural details.

Removing FreqEn leaves more residual degradation and weakens fine structures around object boundaries, indicating that frequency enhancement is beneficial for suppressing degraded components while preserving useful details. Without SFCEM, local structures become less distinct, particularly in the enlarged regions, showing the importance of spatial–frequency interaction for preserving structural information. Removing FMGF further reduces the effective integration of complementary infrared and visible cues, resulting in less prominent target information and weaker local details.

The loss-function variants exhibit complementary effects. Removing $\mathcal{L}_{\mathrm{grad}}$ produces softer object boundaries and less distinct structural details, whereas removing $\mathcal{L}_{\mathrm{color}}$ introduces more noticeable color artifacts around salient regions. In comparison, removing FTCM causes relatively smaller differences in an individual frame because FTCM primarily models temporal dependencies rather than single-frame appearance; its contribution is more directly reflected by the temporal-consistency metrics reported in the main paper.

Overall, the complete FREST preserves clearer object contours, more stable local structures, and a more natural visual appearance, supporting the complementary roles of the proposed components.

## B Computational Efficiency

We further compare the model size, computational cost, and inference time of representative image- and video-level fusion methods at the same input resolution. The results are reported in Table 3.

Table 3: Computational efficiency comparison at an input resolution of $640 \times 480$.

| Method | Pub. | Params (M) | FLOPs (G) | Time (s) |
|---|---|---|---|---|
| LRRNet | *TPAMI'23* | **0.049** | **14.167** | 0.140 |
| CrossFuse | *InfFus'24* | 23.310 | <u>133.706</u> | 0.580 |
| GIFNet | *CVPR'25* | <u>0.823</u> | 226.116 | 0.123 |
| FreeFusion | *TPAMI'25* | 5.670 | 451.655 | <u>0.078</u> |
| TemCoCo | *ICCV'25* | 19.213 | 493.389 | 0.083 |
| UniVF | *NeurIPS'25* | 9.158 | 1509.996 | 0.248 |
| MAVFusion | *ECCV'26* | 9.149 | 169.795 | **0.058** |
| **FREST** | – | 10.887 | 477.788 | 0.255 |

FREST is not the most lightweight model in terms of parameter count, computational cost, or inference time. Nevertheless, its computational requirements remain within a practical range relative to recent video-fusion methods. More importantly, the additional cost supports the joint spatial–frequency representation learning and spatiotemporal frequency modeling that are central to the proposed framework.

Together with the quantitative and qualitative results reported in the main paper, the comparison indicates that FREST provides a favorable performance–complexity trade-off: although it does not minimize model complexity, it achieves stronger overall fusion quality and temporal consistency relative to the compared methods without requiring an excessive increase in computational cost.