AGENTVIF: EVENT-AWARE MULTI-AGENT ADAPTATION TO DYNAMIC DEGRADATIONS IN VISIBLE–INFRARED VIDEO FUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visible-infrared video fusion aims to integrate complementary texture and thermal information into temporally consistent fused videos. However, existing methods commonly rely on fixed fusion parameters and struggle with dynamic degradations whose types and severities vary over time. We propose an event-aware multi-agent framework that couples temporal degradation perception with parameter-space expert composition. A Sentinel Agent localizes degradation transitions, while a customized Expert Agent predicts continuous multi-label degradation severities from dual-modal bursts. These predictions guide a Fusion Mixer to adaptively merge degradation-specific task vectors and generate segment-specific fusion parameters. A Feedback Agent further detects local fusion failures and triggers re-diagnosis and re-fusion. We additionally construct a 20K dual-modal burst-level instruction dataset to customize the Expert Agent. Extensive experiments demonstrate improved fusion quality, temporal consistency, and downstream tracking performance across diverse video conditions.

1 INTRODUCTION

Visible-infrared fusion aims to integrate complementary information from visible and infrared modalities, producing fused results with both rich texture details and reliable thermal responses Ma et al. (2019). This task has been widely studied and applied to various scenarios, such as surveillance Liu et al. (2024a), autonomous driving Li et al. (2023b), and military detection Liu et al. (2025c). Most existing studies focus on image-level fusion Ma et al. (2020); Guan et al. (2026), while visible-infrared video fusion has recently received increasing attention. In video fusion, continuous multi-modal streams require the model to maintain temporal consistency and adapt to time-varying conditions.

However, real-world visible-infrared videos are often affected by dynamic degradations whose types and severities vary over time, such as low-light, over-exposure, haze/fog, motion blur, sensor noise, or unstable modality responses. These degradations may appear, disappear, or change in severity over time due to illumination variation, camera motion, weather changes, or sensor instability Tang et al. (2022). Although existing visible-infrared video fusion methods exploit temporal information to improve sequence consistency, they usually apply a fixed fusion behavior to the whole video Zhao et al. (2025); Li et al. (2025b). Such a static paradigm is insufficient under dynamic degradations, where different temporal segments may require different modality enhancement and fusion strategies. This motivates an adaptive fusion mechanism that can perceive degradation changes over time and adjust the fusion model accordingly.

Recent advances in VLM-based agents provide a promising direction for building adaptive low-level vision systems Zhu et al. (2025a). With visual perception, language reasoning, and decision-making abilities, these agents can interpret complex visual conditions and coordinate specialized models for adaptive processing Zhao et al. (2024). Inspired by this, we explore whether an agentic system can provide adaptive perception and decision-making for visible-infrared video fusion. Existing agentic low-level vision systems mostly follow a tool-selection paradigm, where the agent selects or invokes predefined restoration or fusion tools Li et al. (2025a); Lin et al.. However, for video fusion with continuously changing and compound degradations, such a discrete tool invocation mechanism is not

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.

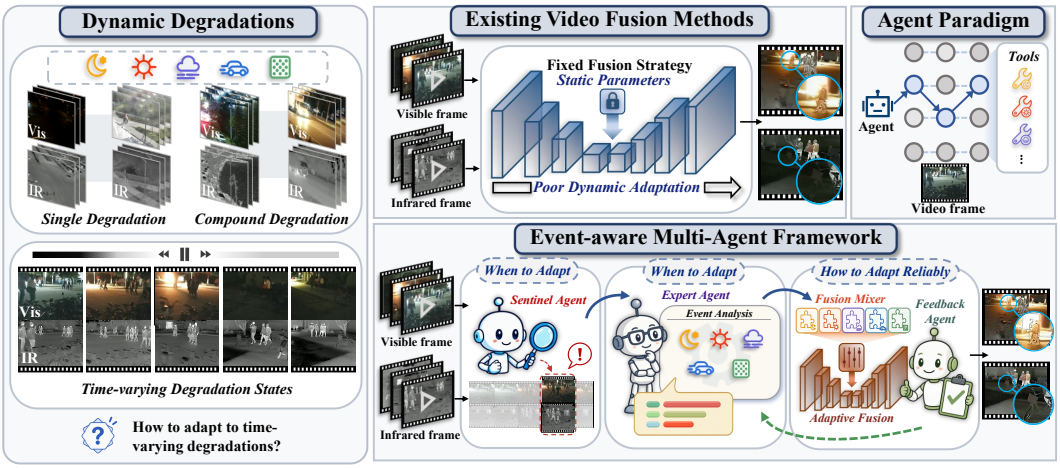

Figure 1: Motivation of AgentVIF for adaptive video fusion.

ideal, since different temporal segments may require fine-grained adaptation rather than selecting a single tool. Moreover, complex degradation conditions often involve multiple factors with different severities, requiring the fusion model to combine different degradation-handling capabilities in a graded manner. These observations motivate an agentic fusion framework that connects temporal degradation perception with parameter-space model adaptation.

To address these challenges, we propose an event-aware multi-agent framework for visible-infrared video fusion, which couples temporal degradation perception with degradation-adaptive expert fusion. The Sentinel Agent first scans the input video pair and localizes sparse degradation transition events. Around each event boundary, the Expert Agent analyzes a short visible-infrared burst and predicts a continuous multi-label degradation severity vector. Instead of using the diagnosis to select a single predefined tool, we introduce an agent-guided Fusion Mixer that dynamically composes degradation-specific fusion experts in the parameter space, producing segment-specific fusion parameters until the next degradation event is detected. Finally, the Feedback Agent checks the fused sequence and triggers local re-diagnosis and re-fusion when potential fusion failures occur. In this way, the agent diagnosis serves as a severity-sensitive control signal for model adaptation, linking event-level degradation perception with parameter-space expert composition for robust visible-infrared video fusion. Our main contributions are as follows:

- We propose an event-aware multi-agent framework for visible-infrared video fusion, where multiple agents collaboratively perform video degradation localization, burst-level diagnosis, and segment-wise adaptive fusion under dynamic degradations.
- We construct a dual-modal burst-level instruction dataset to customize our Expert Agent for continuous multi-label degradation diagnosis, whose estimated severity vector provides severity-aware control signals for adaptive fusion.
- We introduce an agent-guided Fusion Mixer. Unlike tool-selection-based agent systems, it dynamically composes degradation-specific parameters according to the agent diagnosis for adaptive video fusion.

2 Related Work

2.1 Visible-Infrared Video Fusion

Infrared–visible image fusion combines texture from visible cameras with thermal saliency from infrared sensors. Deep models have progressed from representation learning and feature decomposition, such as LRRNet and CDDFuse, to diffusion- and prompt-guided methods, such as DDFM, Mask-DiFuser, and PromptFusion Li et al. (2023a); Zhao et al. (2023b;c); Tang et al. (2025a); Liu et al. (2024a). However, directly applying image fusion frame by frame ignores temporal dependencies and often produces flickering. RCVS jointly handles registration and fusion for video

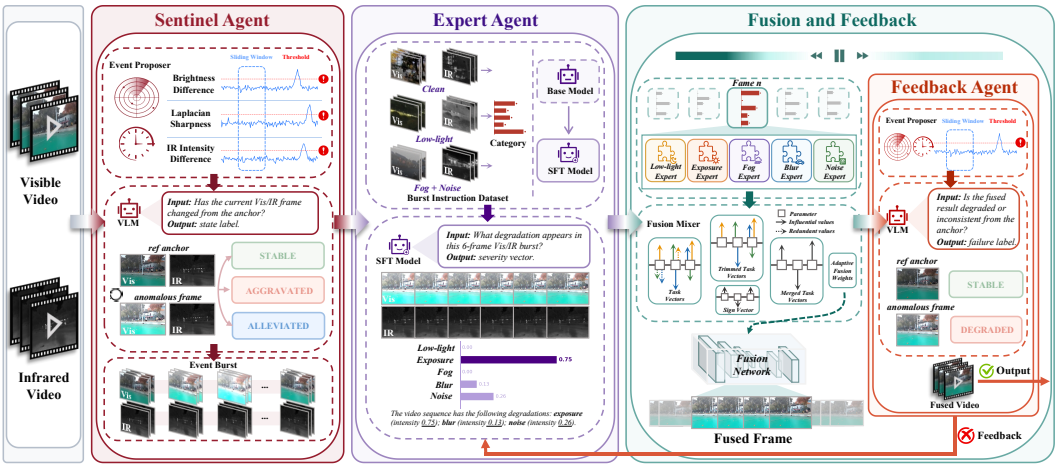

Figure 2: Overall pipeline of AgentVIF.

streams Xie et al. (2024); VideoFusion uses spatio-temporal collaboration for fusion and restoration Tang et al. (2025b); TemCoCo introduces visual-semantic collaboration for temporally consistent fusion Gong et al. (2025); and UniVF develops a unified multi-frame fusion framework Zhao et al. (2026). Although these methods improve temporal coherence, they usually keep one fusion model or strategy throughout a sequence. Consequently, they are difficult to adapt when degradation types and severities change abruptly or differently across modalities.

2.2 Agent-Based Low-Level Vision

Recent advances in LLMs and VLMs have promoted adaptive approaches to low-level vision. Prompt-based methods such as PromptIR and InstructIR condition restoration models on degradation-specific prompts or user instructions Conde et al. (2024); Potlapalli et al. (2023). Beyond prompt conditioning, agent-based systems can perceive degradations, plan suitable operations, invoke expert models, and evaluate intermediate results. For image restoration, AgenticIR improves tool planning through expert experience, reflection, and rollback Zhu et al. (2025a), while HybridAgent coordinates FastAgent, SlowAgent, and FeedbackAgent for efficient restoration Li et al. (2025a). At the video level, MoA-VR employs a mixture of agents for all-in-one video restoration Liu et al. (2025b). These works demonstrate the value of collaboration among specialized agents. Inspired by them, we ask whether a multi-agent framework can be constructed for visible–infrared video fusion, where agents jointly detect temporal degradation events, diagnose local video bursts, adapt fusion behavior, and inspect the outputs.

3 Method

3.1 Overview

We propose a multi-agent visible-infrared video fusion framework that couples temporal degradation perception with adaptive expert fusion. Given a visible-infrared frame sequence $\{(V_t, I_t)\}_{t=1}^{T}$, our goal is to generate a high-quality fused sequence $\mathcal{F} = \{F_t\}_{t=1}^{T}$ under dynamic degradations.

The framework dynamically guides the fusion process through three collaborative agents. The **Sentinel Agent** first parses the input video pair and constructs a degradation event timeline. The **Expert Agent** then estimates the degradation composition of each event segment. Afterwards, the adaptive fusion module performs segment-wise fusion according to the diagnosis results, and refines local segments based on the feedback from the **Feedback Agent**. The pipeline is illustrated in Fig. 2.

3.2 Sentinel Agent

The Sentinel Agent is responsible for localizing temporal transitions of degradation states in visible-infrared videos, including both degradation aggravation and alleviation. Given the input sequence

$\{(V_t, I_t)\}_{t=1}^{T}$, it constructs a sparse event timeline $\mathcal{E} = \{e_k\}_{k=1}^{K}$, where each event indicates a degradation state change with respect to the maintained temporal anchor. The Sentinel Agent only performs event-level localization and extracts a short visible-infrared burst for each event for subsequent diagnosis.

3.2.1 LIGHTWEIGHT EVENT PROPOSAL.

For video sequences, frame-wise VLM judgment is often impractical. Therefore, we use a lightweight CV radar as a temporal event proposer before VLM verification. For adjacent frame pairs (V_{t-1}, I_{t-1}) and (V_t, I_t), this module monitors three complementary inter-frame cues: visible illumination change, visible sharpness change, and infrared response change. These cues are averaged within a short temporal window. When the following condition is satisfied, a candidate transition is proposed and sent to the VLM for semantic verification:

$$\bar{b}_t > \tau_b \quad \text{or} \quad \bar{l}_t > \tau_l \quad \text{or} \quad \bar{r}_t > \tau_r, \tag{1}$$

where $\bar{b}_t$, $\bar{l}_t$, and $\bar{r}_t$ are the window-averaged illumination, sharpness, and infrared response changes, respectively. To reduce missed detections caused by weak or slowly accumulated changes, VLM verification is also triggered every N frames as a heartbeat check.

3.2.2 VLM VERIFICATION.

In the semantic verification stage, we use Qwen3-VL-8B-Instruct with rule-guided contextual prompting to determine whether the current frame undergoes a degradation state transition with respect to the temporal anchor. The visible and infrared frames are horizontally concatenated into a dual-modal image. Given the dual-modal image of the temporal anchor $X_a = [V_a, I_a]$ and that of the current frame $X_t = [V_t, I_t]$, the VLM predicts a relative transition state:

$$\begin{aligned} &s_t = \Phi_{\text{vlm}}(X_a, X_t), \\ &s_t \in \{\texttt{STABLE}, \texttt{AGGRAVATED}, \texttt{ALLEVIATED}\}. \end{aligned} \tag{2}$$

The temporal anchor is initialized as the first frame and updated whenever a new state transition is confirmed. The prompt specifies the dual-modal layout, the degradation transition criteria, and the required JSON output format.

If the VLM outputs AGGRAVATED or ALLEVIATED, the corresponding location is identified as an event boundary s_k. For transitions detected by heartbeat checking, we further perform binary search with VLM verification between the previous anchor and the current frame to refine the boundary. The temporal anchor is then updated to the localized boundary frame, from which a short burst $\mathcal{B}_k = \{(V_t, I_t)\}_{t=s_k}^{s_k+W-1}$ is extracted and used as the input to the Expert Agent.

3.3 EXPERT AGENT

Given the event burst $\mathcal{B}_k$ extracted by the Sentinel Agent, the Expert Agent predicts a continuous multi-label degradation severity vector:

$$\mathbf{p}_k = [p_k^{\text{low}}, p_k^{\text{exp}}, p_k^{\text{fog}}, p_k^{\text{blur}}, p_k^{\text{noise}}] \in [0, 1]^5, \tag{3}$$

corresponding to low-light, over-exposure, fog, blur, and noise. This representation supports compound degradation states with independently varying severities and provides the control signal for adaptive expert fusion.

3.3.1 BURST-LEVEL INSTRUCTION DATASET.

To specialize the Expert Agent for dual-modal degradation diagnosis, we construct a burst-level instruction dataset from temporally aligned visible-infrared video clips. The seed clips are collected from M3SVD Tang et al. (2025b), VT-MOT Zhu et al. (2025b), and NOT156 Sun et al. (2025). Each sample contains six visible frames and six corresponding infrared frames, enabling the model to observe short-term temporal context and cross-modal degradation differences.

We first estimate inherent degradations in the original clips, including low-light, over-exposure, and blur, and use them as initial severity labels. Additional degradation operators covering low-light,

over-exposure, fog, blur, and noise are then applied. The severity of each synthetic degradation is obtained by normalizing its applied strength to $[0, 1]$, while absent degradation types are assigned a score of zero.

Each instruction sample takes twelve images as input, where the first six images are visible frames and the last six images are temporally aligned infrared frames. The target response is organized as a structured diagnosis:

$$\mathbf{y}_k = \{\texttt{probs} : \mathbf{p}_k,\ \texttt{description} : d_k\}, \tag{4}$$

where $\mathbf{p}_k$ denotes the degradation severity vector and d_k provides a concise textual description of the degradation composition. In total, the dataset contains 20K burst-level samples, including 18K samples for training and 2K samples for validation. Among all samples, 3,776 are clean, 5,948 contain a single degradation, and 10,276 contain multiple degradations.

3.3.2 LoRA-based Instruction Tuning.

Based on the constructed instruction dataset, we adapt Qwen3-VL-8B-Instruct into the Expert Agent through LoRA-based supervised instruction tuning. Given a dual-modal burst $\mathcal{B}_k$ and the diagnosis instruction, the model is optimized to generate the structured response $\mathbf{y}_k$. The training objective is formulated as:

$$\mathcal{L}_{\text{sft}} = -\log p_\theta(\mathbf{y}_k \mid \mathcal{B}_k), \tag{5}$$

where $\mathcal{B}_k$ denotes the visible-infrared burst around the k-th event. After adaptation, the Expert Agent learns to identify the degradation composition of each event segment rather than assigning a single degradation category. During inference, it outputs the continuous severity vector $\mathbf{p}_k$, which is passed to the Fusion Mixer for degradation-adaptive expert fusion.

3.4 Agent-guided Fusion and Feedback

The agent-guided fusion stage converts the degradation severity vector predicted by the Expert Agent into adaptive fusion parameters. Given the event-level diagnosis $\mathbf{p}_k$, the Fusion Mixer combines degradation-specific experts to perform segment-wise visible-infrared fusion.

3.4.1 Degradation Experts.

We specialize VideoFusion Tang et al. (2025b) into degradation-specific fusion experts through fine-tuning on controlled degraded visible-infrared video datasets. We adopt VideoFusion as the fusion backbone because it maintains temporal consistency in video fusion while incorporating modality reconstruction branches, which help preserve and enhance modality-specific information under degraded conditions.

We select 60 high-quality visible-infrared video pairs from M3SVD and VT-MOT as reference sequences. For the five degradation types, i.e., low-light, over-exposure, fog, blur, and noise, we apply the corresponding degradation operator to the reference sequences and construct a separate single-degradation dataset for each type. Each dataset contains about $27K$ aligned visible-infrared frame pairs, and all degradation types use the same reference videos to keep expert training comparable.

Starting from the pretrained VideoFusion parameters θ_0, we fine-tune one expert θ_g on each degradation-specific dataset $\mathcal{D}_g$, where $g \in \mathcal{G} = \{\text{low}, \text{exp}, \text{fog}, \text{blur}, \text{noise}\}$. Each expert is represented by its task vector $\Delta_g = \theta_g - \theta_0$, which captures the parameter shift from the base model to the corresponding degradation expert.

3.4.2 Fusion Mixer.

Given the raw degradation severity vector $\mathbf{p}_k$ predicted for the k-th event burst, the Fusion Mixer directly uses this vector as the degradation state of the following temporal segment until the next event is detected. The Fusion Mixer dynamically composes the corresponding degradation experts and applies the resulting segment-specific parameters θ_k to fuse all frames within this segment.

We first activate experts whose severity scores exceed a threshold and normalize their scores within the active set:

$$\alpha_{k,g} = \frac{p_{k,g}}{\sum_{h \in \mathcal{A}_k} p_{k,h} + \epsilon}, \tag{6}$$

where $\mathcal{A}_k$ denotes the active expert set. For each expert task vector Δ_g, small-magnitude updates are removed, resulting in $\widetilde{\Delta}_g$. The dominant update direction is determined by

$$m_{k,j} = \text{sign}\left(\sum_{g\in\mathcal{A}_k} \alpha_{k,g}\widetilde{\Delta}_{g,j}\right). \tag{7}$$

Only sign-consistent updates are retained and merged using their severity weights:

$$\widehat{\Delta}_{k,j} = \frac{\sum_{g\in\mathcal{A}_k} \alpha_{k,g}\mathbb{I}[\text{sign}(\widetilde{\Delta}_{g,j}) = m_{k,j}]\widetilde{\Delta}_{g,j}}{\sum_{g\in\mathcal{A}_k} \alpha_{k,g}\mathbb{I}[\text{sign}(\widetilde{\Delta}_{g,j}) = m_{k,j}] + \epsilon}. \tag{8}$$

To preserve the absolute degradation severity, we further introduce an aggregate adaptation scale:

$$\lambda_k = 1 - \prod_{g\in\mathcal{A}_k}(1 - p_{k,g}), \qquad \theta_k = \theta_0 + \lambda_k\widehat{\Delta}_k. \tag{9}$$

Here, $\alpha_{k,g}$ controls the relative contribution of each expert, while λ_k determines the overall displacement from the base model. This allows the Fusion Mixer to respond to both degradation composition and severity.

3.4.3 Feedback Agent.

After the initial segment-wise fusion, the Feedback Agent scans the fused sequence to detect local fusion failures. It follows the same event-triggered design as the Sentinel Agent: lightweight visual cues first propose suspicious locations, and the VLM then verifies whether the fused frame is degraded.

For each verified failure location, the system extracts the corresponding visible-infrared burst from the original input sequence and sends it back to the Expert Agent for re-diagnosis. The updated severity vector is then used by the Fusion Mixer to recompose the fusion parameters and re-fuse the affected temporal segment. The feedback process stops when no new failure is detected or the maximum number of feedback rounds is reached.

4 Experiments

4.1 Experimental Setup

Dataset Setting. We evaluate our method on the test sequences of **HDO** Xie et al. (2024), **M3SVD** Tang et al. (2025b), and **VT-MOT** Zhu et al. (2025b). These datasets cover diverse real-world scenarios and contain natural degradations such as low-light and over-exposure. In addition, all data used for model customization are drawn from the training splits: the degradation-specific fusion experts are constructed using M3SVD and VT-MOT, while the instruction-tuning data for the Expert Agent are collected from M3SVD, VT-MOT, and NOT-156.

Existing datasets usually contain degradation conditions that remain relatively stable within a video, making it difficult to examine responses to temporal degradation changes. We therefore construct DIVD from 30 spatially aligned visible–infrared videos from HDO, M3SVD, and VT-MOT, totaling 2,987 frame pairs, and introduce five degradation types with varying temporal locations and strengths. Each sequence contains one or more non-overlapping events with sudden or gradual transitions. DIVD is used to analyze degradation-transition localization, diagnosis, and the corresponding fusion response.

Implementation Details All three agents employ Qwen3-VL-8B-Instruct Bai et al. (2025) as their vision-language backbone. Supervised fine-tuning of the Expert Agent is conducted using LoRA within the LLaMA-Factory framework Zheng et al. (2024) for one epoch with a learning rate of 1×10^{-5}. We construct the degradation-specific fusion experts by separately fine-tuning a pretrained VideoFusion model on the corresponding degradation data. Each expert is optimized using AdamW for 20 epochs with a learning rate of 5×10^{-5}. All training and inference experiments are conducted on two NVIDIA GeForce RTX 4090 GPUs.

Table 1: Quantitative comparison of fusion quality on four datasets. The best and second-best results are highlighted with red and blue backgrounds, respectively.

Method	HDO			M3SVD			VT-MOT			DIVD		
	MUSIQ↑	CLIP.↑	SSIM↑	MUSIQ↑	CLIP.↑	SSIM↑	MUSIQ↑	CLIP.↑	SSIM↑	MUSIQ↑	CLIP.↑	SSIM↑
CDDFuse	42.400	0.174	0.610	69.843	0.314	0.620	46.759	0.152	0.642	63.404	0.268	0.623
DDFM	37.470	0.175	0.608	61.593	0.323	0.623	44.379	0.184	0.629	55.304	0.258	0.626
LRRNet	39.606	0.165	0.514	68.347	0.283	0.618	41.301	0.132	0.526	62.152	0.248	0.612
MetaFusion	43.502	0.143	0.482	65.257	0.315	0.568	41.613	0.121	0.486	59.398	0.257	0.570
Mask-DiF	40.897	0.152	0.509	66.136	0.312	0.584	45.469	0.153	0.537	59.807	0.281	0.578
PromptF.	37.046	0.189	**0.615**	64.731	**0.348**	0.624	48.474	0.200	0.6483	59.040	0.300	0.633
TIMFusion	25.878	0.113	0.499	68.384	0.277	0.609	26.509	0.082	0.596	60.703	0.224	0.610
DCEvo	42.106	0.158	0.610	69.514	0.313	0.624	48.481	0.168	0.644	62.897	0.268	0.635
GIFNet	43.410	0.204	0.506	66.796	0.264	0.621	43.723	0.187	0.526	62.723	0.242	0.610
SAGE	31.787	0.126	0.538	69.555	0.312	**0.630**	36.741	0.135	0.556	61.243	0.290	**0.638**
TemCoCo	42.941	0.214	0.546	69.878	0.293	0.591	40.013	0.153	0.581	63.410	0.254	0.603
UniVF	34.471	0.191	0.607	63.043	0.269	0.618	42.388	0.147	0.642	56.437	0.225	0.632
VideoFusion	37.652	0.264	0.563	68.259	0.277	0.624	40.179	0.187	0.596	62.697	0.301	0.631
Ours	**43.581**	**0.306**	0.611	**69.998**	0.338	0.627	**48.597**	**0.254**	**0.650**	**63.507**	**0.310**	**0.638**

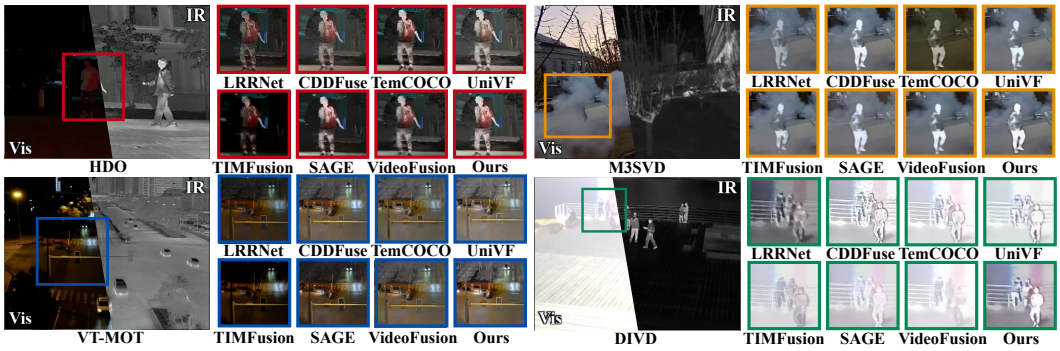

Figure 3: Qualitative comparisons between our method and existing fusion approaches.

4.2 Results of Multi-Modality Video Fusion

In this section, we compare our method with 13 representative fusion methods, including CDDFuse Zhao et al. (2023b), DDFM Zhao et al. (2023c), LRRNet Li et al. (2023a), MetaFusion Zhao et al. (2023a), Mask-DiF Tang et al. (2025a), PromptFusion Liu et al. (2024b), TIMFusion Liu et al. (2024b), DCEvo Liu et al. (2025a), GIFNet Cheng et al. (2025), and SAGE Wu et al. (2025), and video-based methods: TemCoCo Gong et al. (2025), UniVF Zhao et al. (2025), and VideoFusion Tang et al. (2025b).

Quantitative Comparisons. Since conventional fusion metrics can be sensitive to source-image quality and spatial alignment under real-world degradations, we combine no-reference perceptual assessment with structural fidelity evaluation. Specifically, MUSIQ Ke et al. (2021) and CLIP-IQA Wang et al. (2023) are employed to measure overall perceptual quality and semantic fidelity, while SSIM Wang et al. (2004) evaluates the preservation of structural information from the source modalities. As shown in Table 1, our method ranks first or second in all 12 dataset–metric combinations and achieves the best results in nine of them. Specifically, it obtains the best MUSIQ and CLIP-IQA scores on HDO and the best MUSIQ score on M3SVD. More notably, our method ranks first across all three metrics on both VT-MOT and DIVD. These results demonstrate the robustness of our method under temporally varying degradation conditions.

Qualitative Comparisons. As shown in Fig. 3, our method consistently produces clearer and more natural fusion results across the four datasets. On HDO and M3SVD, it preserves visible details

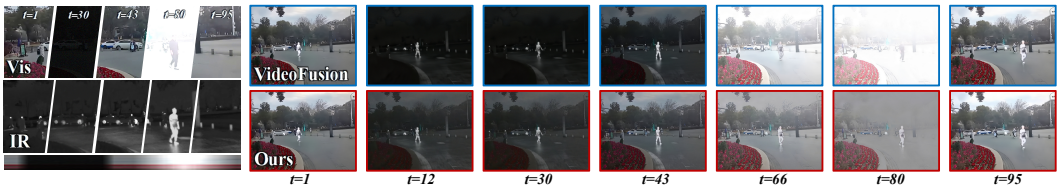

Figure 4: Temporal fusion comparison under dynamically varying degradations.

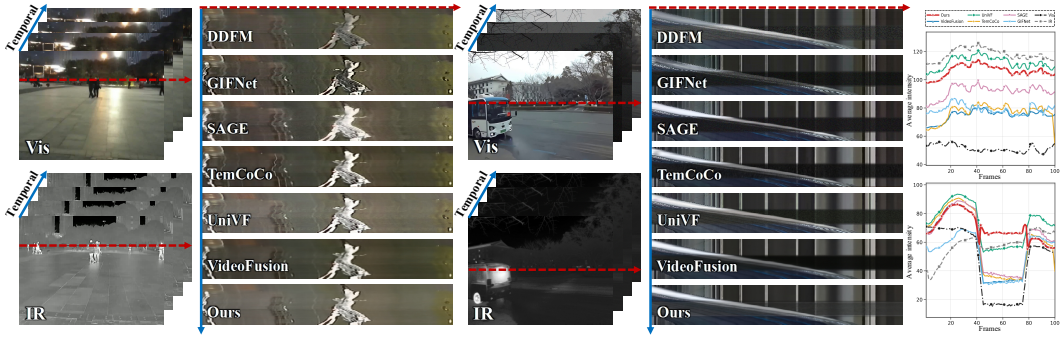

Figure 5: Temporal comparison on VT-MOT and DIVD. Temporal profiles and frame-wise average intensity curves are shown for different fusion methods.

while maintaining salient infrared targets under low-light and low-contrast conditions. On VT-MOT, vehicle structures and road details are clearly retained without amplifying noise. For the overexposed DIVD scene, our method alleviates highlight saturation and preserves clearer pedestrian contours than competing methods. Overall, our method achieves a favorable balance between visible textures and infrared salient information under diverse degradations. Fig. 4 further shows that our method maintains more stable brightness and clearer structural details than VideoFusion as the degradation state changes over time.

4.3 Temporal Consistency Comparison

Following Pan et al. Pan et al. (2023), we visualize the frame-wise luminance variations of sequences from VT-MOT and DIVD in Fig. 5. On VT-MOT, our method produces a smoother luminance curve with fewer abrupt fluctuations. In the DIVD sequence, the later frames gradually transition to a low-light condition, causing the outputs of competing methods to become noticeably darker. In contrast, our method maintains relatively stable brightness and clear scene visibility throughout the sequence, demonstrating better temporal consistency and adaptability to dynamically changing degradations.

4.4 Ablation Study

To evaluate individual agents and the expert-merging strategy, we conduct five ablations. Specifically, **(a) w/o Agent 1** replaces event-aware localization with uniform sampling of six visible-infrared frame pairs; **(b) w/o Agent 2 tuning** uses the original zero-shot VLM instead of the tuned Expert Agent; **(c) w/o Agent 3** removes fusion-quality feedback; **(d) Max-weight selection** retains only the expert with the highest degradation score; and **(e) Direct weighted merging** directly computes a severity-weighted average of expert task vectors using normalized degradation scores, without magnitude pruning or sign-conflict resolution. **(f)** denotes our complete model.

As shown in Fig. 6, the complete model achieves the best overall performance on VT-MOT. Fig. 7 shows that the complete model produces more temporally stable results as the degradation condition varies in DIVD. The qualitative results in Fig. 8 further demonstrate better detail preservation and degradation suppression on VT-MOT and DIVD.

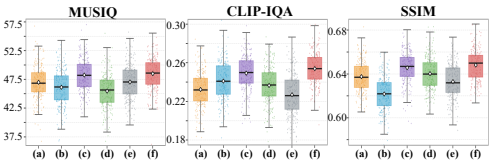

Figure 6: Ablation results on VT-MOT.

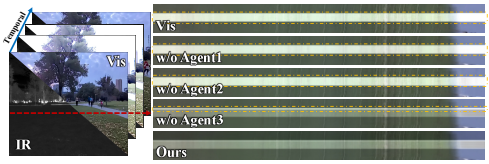

Figure 7: Temporal consistency comparison of agent ablations on DIVD.

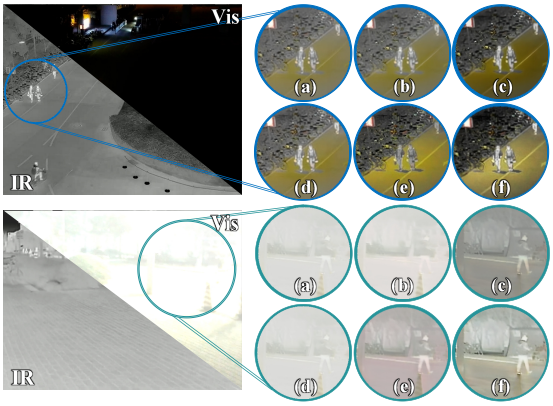

Figure 8: Qualitative ablation comparisons on VT-MOT and DIVD.

Table 2: Quantitative comparison of object tracking performance.

Method	MOTA ↑	MOTP ↑	AUC ↑	DP@20 ↑
CDDFuse	0.025	0.693	0.219	0.336
DDFM	0.011	0.695	0.213	0.326
LRRNet	0.034	0.694	0.199	0.299
MetaFusion	-0.456	0.567	0.029	0.035
Mask-Dif	0.040	0.691	0.213	0.325
PromptF.	-0.022	0.592	0.122	0.195
TIMFusion	0.035	0.684	0.160	0.237
DCEVO	0.031	0.698	0.215	0.329
GIFNet	0.007	0.696	0.197	0.304
SAGE	0.035	0.699	0.181	0.270
TemCOCO	0.048	0.695	0.187	0.290
UniVF	0.043	0.697	0.225	0.348
VideoFusion	0.049	0.696	0.205	0.313
Ours	**0.056**	**0.708**	**0.237**	**0.360**

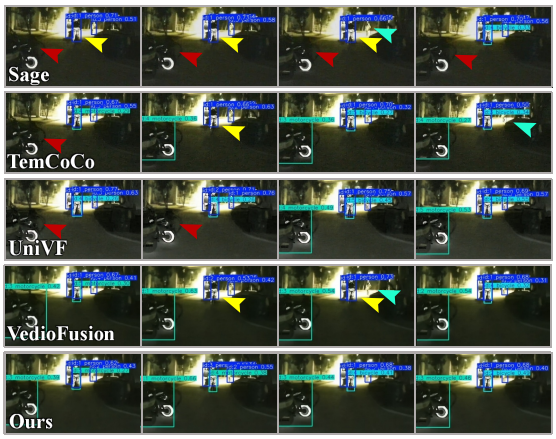

Figure 9: Visual comparison of object tracking.

4.5 Object Tracking

To evaluate the downstream utility of the fused videos, we conduct object tracking experiments on the VT-MOT dataset. We use ByteTrack with a pre-trained YOLOv11n detector without fine-tuning, and report MOTA, MOTP, AUC, and Distance Precision at 20 pixels (DP@20). As shown in Table 2, our method achieves the best performance across all four tracking metrics. These results demonstrate that our method can provide stable and reliable fused results for object tracking under dynamic degradation scenarios.

As shown in Fig. 9, competing fusion methods exhibit missed detections or unstable target localization in challenging scenes. In contrast, our method maintains more continuous and accurate target responses across consecutive frames, further demonstrating its reliability for downstream video perception under dynamic degradations.

5 Conclusion

In this work, we proposed an event-aware multi-agent framework for visible-infrared video fusion under dynamic degradations. It localizes degradation transitions, diagnoses continuous multi-label severities from dual-modal bursts, and composes degradation experts into segment-specific fusion models. A feedback mechanism further corrects local fusion failures. We also constructed a burst-level instruction dataset to customize degradation diagnosis. Experiments on four datasets demonstrate improved fusion quality, temporal consistency, and downstream tracking performance, supporting agent-based degradation reasoning as an effective control signal for adaptive video fusion.

AI USE STATEMENT

Generative AI tools were used during manuscript preparation to improve textual clarity and organization, refine scientific figures, and provide critical feedback on the presentation of the methodology and experimental analysis. The authors manually examined and revised all resulting suggestions against the original experiments, data, and cited sources. All scientific decisions and conclusions remain those of the authors, who assume full responsibility for the submitted work.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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 Computer Vision and Pattern Recognition Conference*, pp. 28102–28112, 2025.

Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *European Conference on Computer Vision*, pp. 1–21. Springer, 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.

Tianwei Guan, Haozhen Wei, Yuhan Zhou, Jun Ma, Zecheng Xu, Zhiying Jiang, Jinyuan Liu, and Xingyuan Li. Domain adaptation guided infrared and visible image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4376–4384, 2026.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Bingchen Li, Xin Li, Yiting Lu, and Zhibo Chen. Hybrid agents for image restoration. *arXiv preprint arXiv:2503.10120*, 2025a.

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, 2023a.

Xingyuan Li, Yang Zou, Jinyuan Liu, Zhiying Jiang, Long Ma, Xin Fan, and Risheng Liu. From text to pixels: A context-aware semantic synergy solution for infrared and visible image fusion. *arXiv preprint arXiv:2401.00421*, 2023b.

Zhuoyuan Li, Junqi Liao, Chuanbo Tang, Haotian Zhang, Yuqi Li, Yifan Bian, Xihua Sheng, Xinmin Feng, Yao Li, Changsheng Gao, et al. Ustc-td: A test dataset and benchmark for image and video coding in 2020s. *IEEE Transactions on Multimedia*, 2025b.

ZiXu Lin, Yunlong Lin, Kunjie Lin, Kepeng Yang, Ge Meng, Yingying Wang, Yue Huang, and Xinghao Ding. Fuseagent: A vlm-driven agent for unified in-the-wild image fusion.

Jinyuan Liu, Xingyuan Li, Zirui Wang, Zhiying Jiang, Wei Zhong, Wei Fan, and Bin Xu. Promptfusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*, 2024a.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2226–2235, 2025a.

Lu Liu, Chunlei Cai, Shaocheng Shen, Jianfeng Liang, Weimin Ouyang, Tianxiao Ye, Jian Mao, Huiyu Duan, Jiangchao Yao, Xiaoyun Zhang, et al. Moa-vr: A mixture-of-agents system towards all-in-one video restoration. *IEEE Journal of Selected Topics in Signal Processing*, 2025b.

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

Yating Liu, Yang Zou, Xingyuan Li, Xingyue Zhu, Kaiqi Han, Zhiying Jiang, Long Ma, and Jinyuan Liu. Toward a training-free plug-and-play refinement framework for infrared and visible image registration and fusion. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1268–1277, 2025c.

Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information fusion*, 45:153–178, 2019.

Jiayi Ma, Hao Zhang, Zhenfeng Shao, Pengwei Liang, and Han Xu. Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Transactions on Instrumentation and Measurement*, 70:1–14, 2020.

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

Vaishnav Potlapalli, Syed Waqas Zamir, Salman H Khan, and Fahad Shahbaz Khan. Promptir: Prompting for all-in-one image restoration. *Advances in neural information processing systems*, 36:71275–71293, 2023.

Chen Sun, Xinyu Wang, Shenghua Fan, Xiaobing Dai, Yuting Wan, Xiao Jiang, Zengliang Zhu, and Yanfei Zhong. Not-156: Night object tracking using low-light and thermal infrared: From multi-modal common-aperture camera to benchmark datasets. *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83: 79–92, 2022.

Linfeng Tang, Chunyu Li, and Jiayi Ma. Mask-difuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Linfeng Tang, Yeda Wang, Meiqi Gong, Zizhuo Li, Yuxin Deng, Xunpeng Yi, Chunyu Li, Han Xu, Hao Zhang, and Jiayi Ma. Videofusion: A spatio-temporal collaborative network for multi-modal video fusion and restoration. *arXiv preprint arXiv:2503.23359*, 2025b.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 2555–2563, 2023.

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.

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

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

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.

Zhonghan Zhao, Wenhao Chai, Xuan Wang, Boyi Li, Shengyu Hao, Shidong Cao, Tian Ye, and Gaoang Wang. See and think: Embodied agent in virtual environment. In *European Conference on Computer Vision*, pp. 187–204. Springer, 2024.

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. *arXiv preprint arXiv:2505.19858*, 2025.

Zixiang Zhao, Haowen Bai, Bingxin Ke, Yukun Cui, Lilun Deng, Yulun Zhang, Kai Zhang, and Konrad Schindler. A unified solution to video fusion: From multi-frame learning to benchmarking. *Advances in Neural Information Processing Systems*, 38:128504–128535, 2026.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL http://arxiv.org/abs/2403.13372.

Kaiwen Zhu, Jinjin Gu, Zhiyuan You, Yu Qiao, and Chao Dong. An intelligent agentic system for complex image restoration problems. In *International Conference on Learning Representations*, volume 2025, pp. 57985–58013, 2025a.

Yabin Zhu, Qianwu Wang, Chenglong Li, Jin Tang, Chengjie Gu, and Zhixiang Huang. Visible–thermal multiple object tracking: Large-scale video dataset and progressive fusion approach. *Pattern Recognition*, 161:111330, 2025b.

A Details of the Multi-Agent Framework

In this section, we provide additional implementation details of the multi-agent framework introduced in the main paper, including the event proposal mechanisms, agent prompts, input and output formats, and fine-tuning configuration.

A.1 Sentinel Agent

The Sentinel Agent uses a lightweight CV radar as a temporal event proposer before VLM verification. For adjacent frame pairs (V_{t-1}, I_{t-1}) and (V_t, I_t), it monitors three complementary inter-frame cues: visible illumination change, visible sharpness change, and infrared response change.

Let Y_t and Z_t denote the grayscale visible and infrared frames, respectively. The three inter-frame cues are computed as

$$b_t = \frac{1}{|\Omega|} \sum_{x \in \Omega} \left| \mathcal{G}_7(Y_t)(x) - \mathcal{G}_7(Y_{t-1})(x) \right|, \tag{10}$$

$$l_t = \frac{\left| \mathrm{Var}[\mathcal{L}(Y_t)] - \mathrm{Var}[\mathcal{L}(Y_{t-1})] \right|}{\mathrm{Var}[\mathcal{L}(Y_{t-1})] + \epsilon_{\mathrm{cv}}}, \tag{11}$$

$$r_t = \frac{1}{|\Omega_{\mathrm{IR}}|} \sum_{x \in \Omega_{\mathrm{IR}}} \left| Z_t(x) - Z_{t-1}(x) \right|, \tag{12}$$

where $\mathcal{G}_7$ denotes Gaussian filtering with a 7×7 kernel, $\mathcal{L}$ is the discrete Laplacian operator, and $\epsilon_{\mathrm{cv}} = 10^{-5}$. Each cue is averaged over a three-frame temporal window, and a candidate transition is proposed if $\bar{b}_t > \tau_b$, $\bar{l}_t > \tau_l$, or $\bar{r}_t > \tau_r$. We set $(\tau_b, \tau_l, \tau_r) = (12.0, 0.5, 10.0)$. VLM verification is additionally triggered every $N = 10$ frames as a heartbeat check to capture weak or gradually accumulated changes. The complete rule-guided contextual prompt used by the Sentinel Agent is provided in Table 5.

A.2 Expert Agent

The Expert Agent produces a structured diagnosis $y_k = \{\texttt{probs} : \mathbf{p}_k, \texttt{description} : d_k\}$, where $\mathbf{p}_k$ contains the estimated severities of low-light, over-exposure, fog, blur, and noise. Although the numerical are stored under the JSON key `probs`, they represent independent degradation severity scores and are not constrained to sum to one, allowing multiple degradation types to be represented simultaneously. The complete instruction prompt and structured output format are provided in Table 6.

We customize Qwen3-VL-8B-Instruct through LoRA-based supervised instruction tuning using LLaMA-Factory. LoRA adapters are applied to the query, key, value, and output projections in the attention modules, as well as the gate, up, and down projections in the feed-forward modules. We set the LoRA rank, scaling parameter, and dropout rate to 32, 64, and 0.1, respectively. The model is trained for one epoch using paged AdamW with a learning rate of 1×10^{-5}, a cosine learning-rate schedule, and 100 warmup steps. We use bfloat16 precision, a per-device batch size of one, and eight gradient accumulation steps on two GPUs, resulting in an effective batch size of 16. The image pixel budget is set to 100,352 pixels per image.

A.3 Feedback Agent

The Feedback Agent follows the same event-triggered design as the Sentinel Agent, using lightweight visual cues to propose suspicious locations before VLM verification. For the grayscale fused frame Y_t^F, the inter-frame intensity and global sharpness changes are computed in the same manner as b_t and l_t, respectively. To capture spatially localized fusion artifacts, we additionally divide the fused frame into non-overlapping 32×32 blocks and compute

$$u_t = \mathrm{Std}_{P \in \mathcal{P}} \left[\mathrm{Var} \left[\mathcal{L}\big(Y_t^F|_P\big) \right] \right], \qquad c_t = \frac{|u_t - u_{t-1}|}{u_{t-1} + \epsilon_{\mathrm{cv}}}, \tag{13}$$

where $\mathcal{P}$ denotes the set of image blocks and $\epsilon_{\mathrm{cv}} = 10^{-5}$. The cue c_t measures temporal variation in the spatial distribution of local sharpness and is used to propose potential local fusion failures.

All three visual cues are averaged over a three-frame temporal window. The thresholds for inter-frame intensity change, global sharpness change, and local structural variation are set to 10.0, 0.4, and 0.3, respectively. VLM verification is additionally triggered every $N = 10$ frames as a heartbeat check. The complete rule-guided contextual prompt used to verify whether a fused frame is `STABLE` or `DEGRADED` is provided in Table 7.

For each verified fusion failure, the corresponding six-frame visible–infrared burst is sent to the Expert Agent for re-diagnosis, and the affected temporal segment is subsequently re-fused. We perform at most two feedback rounds. In the r-th round, the three proposal thresholds are multiplied by 1.5^{r-1}, and the feedback process terminates early if no new fusion failure is detected.

B Dataset Construction and Split Protocol

This section describes the data used to customize the Expert Agent and the degradation-specific fusion experts, together with the construction of DIVD.

B.1 Data Sources and Separation

We use four aligned visible–infrared video datasets. M3SVD and VT-MOT provide diverse sequences for both model customization and evaluation, but their source videos are partitioned between these uses. Their training-only videos supply the 20K instruction dataset and the five fusion-expert datasets, whereas held-out videos are used for native evaluation and DIVD. NOT-156 is used only to enrich the instruction data with nighttime low-light–thermal sequences, while HDO is reserved for evaluation and DIVD. Thus, although M3SVD and VT-MOT appear in multiple branches, the corresponding source-video identities are disjoint across the customization and evaluation pools.

All partitions are fixed by canonical source-video ID before burst extraction or degradation synthesis. Every burst, crop, and degraded copy inherits the assignment of its source video, preventing adjacent bursts or alternative degradation versions of the same scene from crossing partitions. The two customization branches may share videos within the training-only pool, but neither shares a source video with the evaluation data.

B.2 Burst-Level Instruction Dataset

The instruction dataset is constructed from the training portions of M3SVD, VT-MOT, and NOT-156. After source-level partitioning, temporally contiguous six-frame visible–infrared bursts are sampled across each video, with a per-video cap to avoid domination by long sequences.

We retain naturally occurring low-light, over-exposure, and blur as initial labels. They are estimated from the middle visible frame using its mean intensity and Laplacian variance: mean intensities below 70 and above 180 indicate low-light and over-exposure, respectively, while a Laplacian variance below 100 provides a blur prior. Intensity-based estimation is disabled for near-grayscale visible streams. Synthetic components are produced by the five modality-aware operators described below. Their peak strengths are modulated within the burst by $h_i = \sin(\pi i/5)$, $i \in \{0, \dots, 5\}$, to produce temporally coherent changes.

For degradation g, the native and synthetic components are combined as $c_g = [c_g^{\mathrm{nat}} + s_g]_0^1$. The common severity mapping used for training and evaluation is

$$p_g = \rho(c_g), \qquad \rho(c) = \begin{cases} 0, & c < 0.05, \\ \min\left(1, \dfrac{0.9}{0.7} c\right), & c \geq 0.05. \end{cases} \tag{14}$$

Each entry represents the condition-level severity of the paired burst and is independent of the other entries. Differences between the two modalities are expressed by the operator responses rather than separate modality-specific labels.

The resulting dataset contains 20,000 bursts, with 18,000 used for training and 2,000 for validation. It comprises 3,776 clean, 5,948 single-degradation, and 10,276 compound-degradation samples.

The same operators are shared by the instruction data, fusion-expert data, and DIVD. Let $s \in [0, 1]$ denote the sampled severity, with image intensities normalized to $[0, 1]$.

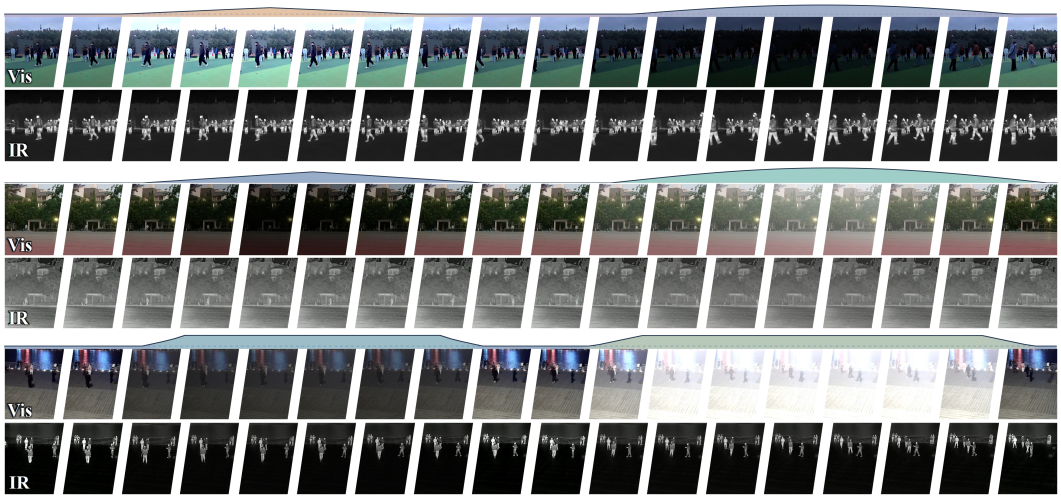

Figure 10: Examples from DIVD. The curves show the applied degradation levels over time, and the visible–infrared frame pairs illustrate the corresponding visual changes.

- **Low-light.** The visible frame is attenuated and gamma-darkened, followed by signal-dependent perturbation. The attenuation factor and gamma are $1 - 0.8s$ and $1 + 1.5s$, respectively, and the noise scale is $0.05s$. Since thermal observations are less sensitive to ambient illumination, the infrared frame retains its mean response and receives only a weak perturbation with scale $0.005s$.
- **Over-exposure.** We combine gain amplification, inverse-gamma mapping, and highlight clipping in the visible modality, using gain $1 + 1.5s$ and gamma $1 - 0.5s$. The infrared modality undergoes only a small gain adjustment of $1 + 0.1s$, preserving most thermal structures.
- **Fog.** Visible fog follows the atmospheric scattering model $V' = T_s \odot V + A(1 - T_s)$, where $A \sim \mathcal{U}(0.85, 0.95)$ and T_s is a spatially smooth transmission field whose range decreases toward $[1 - 0.9s,\ 1 - 0.2s]$. For infrared, we model partial thermal penetration through fog by applying mild contrast attenuation with factor $1 - 0.2s$.
- **Blur.** Both modalities are Gaussian-smoothed, but a smaller kernel is used for infrared to retain salient thermal structures. The kernel sizes are $k_V = \mathrm{odd}(3 + 12s)$ and $k_I = \max(3, k_V - 4)$.
- **Noise.** Visible sensor noise is modeled by a Poisson–Gaussian process whose photon scale decreases as $200 - 150s$, together with read noise of scale $5s/255$. Infrared noise is represented by column-wise fixed-pattern perturbations with standard deviation $12s$ on the 8-bit intensity scale.

For compound samples, the active operators are composed in a fixed physically motivated order: atmospheric scattering, illumination response, sensor noise, and spatial blur. This shared construction keeps the degradation semantics consistent across data branches while allowing their composition and temporal evolution to vary.

B.3 Degradation-Specific Fusion-Expert Data

We select 60 high-quality training videos from M3SVD and VT-MOT and reuse the same video identities, frame ranges, and train–validation assignment for all five experts. For each degradation type, only its corresponding operator is applied, producing a matched single-degradation dataset of approximately 27K aligned frame pairs; the original streams serve as reconstruction targets. This controls scene content across experts so that their task-vector differences mainly reflect degradation-specific adaptation. Training samples use six-frame windows and 128×128 spatial crops.

Table 3: Expert Agent on DIVD. Recognition and frame-level metrics use controlled temporal annotations, while event-level metrics use 135 six-frame windows. Δ denotes the absolute change from the zero-shot backbone; Acc@10 is reported in percentage points (pp).

	Recognition			Frame severity	Event severity	
Method	Macro-F1 ↑	mAP ↑	Micro-F1 ↑	MAE ↓	Acc@10 ↑	MAE ↓
Qwen3-VL-8B-Instruct	0.444	0.556	0.466	0.142	22.41%	0.389
Expert Agent	**0.633**	**0.679**	**0.585**	**0.114**	**32.55%**	**0.262**
Δ vs. zero-shot	$+0.189$	$+0.123$	$+0.119$	-0.028	$+10.14$ pp	-0.127

Table 4: Sentinel Agent on DIVD.. ER: Event Recall; OR@10: Onset Recall within ten frames; IAR: In-event Alert Ratio; DA: Direction Accuracy. Delay is reported as mean/median in frames.

Method	ER ↑	OR@10 ↑	IAR ↑	Delay ↓	DA ↑
Sentinel Agent	**95.45%**	**81.82%**	**93.24%**	**5.69 / 4**	**90.37%**
Counts	*42/44*	*36/44*	*138/148*	–	*122/135*

B.4 Construction of DIVD

DIVD is used to examine how the proposed agents respond when the degradation condition changes over time. It is constructed from 30 aligned videos: 10 from HDO, 13 from M3SVD, and 7 from VT-MOT, totaling 2,987 frame pairs. The selected clips contain no pronounced target degradation, while mild native imperfections are retained as background conditions. We apply the degradation operators described above with time-varying strengths, allowing the applied degradation state to be tracked throughout each sequence. These records are used to compare the events and severity scores predicted by the agents with the imposed degradation changes.

Each sequence contains one or more temporally non-overlapping degradation events. An event specifies its active degradation types, temporal range $[s_e, e_e]$, peak strengths $a_{e,g}$, and temporal envelope $h_e(t)$. The frame-level operator strength and severity label are

$$c_{t,g} = [c_g^{\mathrm{nat}} + a_{e,g} h_e(t)]_0^1, \qquad p_{t,g} = \rho(c_{t,g}), \qquad t \in [s_e, e_e]. \tag{15}$$

DIVD includes rapid transitions, gradual changes, smooth pulses, and plateaued intervals, as well as single and compound events. Each annotation records the event boundary, degradation composition, peak severity, temporal pattern, ramp length, and random seed.

C Agent Capability Evaluation

C.1 Expert Agent.

We evaluate the Expert Agent on 30 DIVD sequences comprising 2,987 frames. For evaluation only, we define the predicted and ground-truth severity changes as

$$\Delta\widehat{p}_{k,g} = [\widehat{p}_{k,g} - \widehat{p}_{0,g}]_+, \qquad \Delta p_{t,g} = [p_{t,g} - p_{0,g}]_+. \tag{16}$$

These baseline-corrected changes are used only for DIVD evaluation; the Fusion Mixer uses the raw severity vector $\widehat{\mathbf{p}}_k$. Event-level predictions are assigned to their corresponding temporal segments. For recognition, predictions and annotations are binarized using the Fusion Mixer's activation threshold $\tau = 0.05$. Macro-F1 and Micro-F1 evaluate degradation recognition, mAP evaluates score ranking, and frame-level MAE measures severity estimation.

Event-level metrics are computed on $N = 135$ six-frame windows. Let Let $\bar{s}_{i,g}$ and $\hat{s}_{i,g}$ denote the mean ground-truth change $\Delta p_{t,g}$ and the corresponding predicted change $\Delta\widehat{p}_{k,g}$ in window i, respectively. With the active set

$$\mathcal{G}_i = \{g \mid \bar{s}_{i,g} > \tau\}, \tag{17}$$

the event error and the two event-level metrics are defined as

$$e_i = \frac{1}{|\mathcal{G}_i|} \sum_{g \in \mathcal{G}_i} |\hat{s}_{i,g} - \bar{s}_{i,g}|,$$
$$\text{EventMAE} = \frac{1}{N} \sum_{i=1}^{N} e_i, \qquad \text{Acc@10} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(e_i \leq 0.10). \tag{18}$$

As shown in Table 3, instruction tuning consistently improves all recognition and severity metrics. The Expert Agent improves Macro-F1, mAP, and Micro-F1 by 0.189, 0.123, and 0.119, respectively, while reducing frame-level and event-level MAE by 0.028 and 0.127. Event-level Acc@10 also increases by 10.14 percentage points. These results show that instruction tuning improves the recognition and severity estimation of controlled temporal degradation changes. We interpret these scores as coarse adaptation strengths rather than precise physical measurements.

C.2 Sentinel Agent.

We evaluate the Sentinel Agent on 44 annotated degradation events using detections from the initial fusion pass. An event is detected if an alert falls within its annotated interval. OR@10 measures whether the first in-event alert occurs within ten frames of the event onset, while IAR measures the proportion of alerts located inside annotated event intervals. Direction accuracy excludes severity changes no larger than 0.05.

Table 4 shows that the Sentinel Agent detects 42 of 44 events, including 36 within ten frames of onset. Moreover, 138 of 148 alerts fall inside annotated event intervals, and 122 of 135 valid transitions have the correct direction. With a median delay of four frames, the Sentinel Agent provides reliable and timely triggers for subsequent degradation diagnosis.

Table 5: Prompt for the Sentinel Agent.

Sentinel Agent

Role. You are an agent for detecting abrupt anomalies in dual-modal videos. Your sole task is to act as an initial sentinel: identify the frame at which environmental degradation changes, either worsening or improving, and trigger an alert. You do not need to identify the degradation type.

Input. Each input image is horizontally concatenated: the left half is the visible image (VI), and the right half is the infrared image (IR). Image 1 is the past reference frame, and Image 2 is the current frame.

Degradation cues. Watch for the following signs of deterioration, occurring individually or in combination:

- Fog or whitish obscuration, particularly in distant visible regions.
- Darkening or overexposed white regions, primarily in the visible half.
- Severe speckle-like noise in either modality.
- Evident compression-related blur or defocus in either modality.

Important multimodal assessment rule. The infrared half can see through fog, so a clear infrared image during fog is normal. If the visible half is clearly foggy, report the change; do not dismiss visible fog simply because the infrared half is clear.

Decision logic. Ignore normal vehicle and pedestrian motion. Focus on environmental image quality and changes in degradation.

- **STABLE**: The overall degradation level is approximately unchanged. Even if both images are dark or foggy, return STABLE when the current image is neither worse nor better than the reference.
- **AGGRAVATED**: The current image is clearly worse than the reference, for example darker, foggier, blurrier, or noisier.
- **ALLEVIATED**: The current image is clearly better than the reference, for example fog has dispersed, illumination has improved, or details have become clearer.

Output format. Return only a JSON object with the key `"status"` and exactly one of `"STABLE"`, `"AGGRAVATED"`, or `"ALLEVIATED"`. Do not output degradation types.

Table 6: Prompt for the Expert Agent.

Expert Agent

Input. The twelve images above form a dual-modal video clip. The first six images are visible (RGB) frames, and the last six are the corresponding infrared (IR) frames. The two groups are strictly aligned in time.

Task. Act as a professional agent for low-level visual quality diagnosis. Jointly analyze the image quality of both modalities and assess the global degradation state of this video clip.

Output format. Your output must strictly follow the JSON format below, without any additional explanation or text:

```
{
  "probs": {
    "lowlight": 0.0,
    "exposure": 0.0,
    "fog": 0.0,
    "blur": 0.0,
    "noise": 0.0
  },
  "description": "..."
}
```

Field definitions.

- `lowlight`: under-exposure or low-light conditions.
- `exposure`: over-exposure.
- `fog`: haze or scattering.
- `blur`: motion blur or defocus blur.
- `noise`: sensor noise.

Each numerical field must lie in $[0.0, 1.0]$, where 0.0 indicates no degradation and 1.0 indicates extremely severe degradation. Use one sentence in `"description"` to summarize the overall degradation.

Table 7: Prompt for the Feedback Agent.

Feedback Agent

Role. You are a fused-video quality inspection agent specializing in quality problems in the outputs of visible–infrared fusion networks.

Input. Image 1 is the reference fused frame. Image 2 is the current fused frame.

Fusion failure modes. Report `DEGRADED` when any of the following fusion failure modes occurs:

1. **Excessive smoothing or detail removal:** textures disappear across substantial regions and edges are smoothed, as if a strong filter had been applied.
2. **Abnormal brightness:** the image becomes substantially darker with large black regions, or overexposed with large white regions.
3. **Amplified noise:** speckle-like noise, block artifacts, or substantially increased graininess appears.
4. **Color shifts:** the overall appearance becomes clearly warmer, cooler, or grayer and differs markedly from the reference.
5. **Structural ghosting:** moving objects exhibit double images, blurred ghost contours, or visibly misaligned overlapping positions.

Do not raise an alarm for:

- Normal vehicle or pedestrian motion. Ghosting means that the same object exhibits two semi-transparent contours.
- A scene change with completely different background content.
- Small global brightness differences within approximately $\pm 10\%$.

Decision rule. Use a sensitive criterion that prioritizes detecting failures over avoiding false alarms.

- **`DEGRADED`:** the current frame exhibits any failure mode listed above, including a moderately pronounced one.
- **`STABLE`:** fusion quality is comparable to the reference and no obvious fusion artifact is present.

Output format. Return only a JSON object with the key `"status"` and exactly one of `"STABLE"` or `"DEGRADED"`. Do not output any additional content.