V2VFusion: Text-Controlled Video-to-Video Diffusion for Degradation-Aware Video Fusion

Jiajun Chen[1], **Han Xu**[1*], **Yunfei Hang**[1], **Zhuoya Zhang**[1], **Guangcan Liu**[1], **Jiayi Ma**[2*]
[1] School of Automation, Southeast University, Nanjing, China
[2] School of Robotics, Wuhan University, Wuhan, China
{jiajunchen,xu_han,huang_yunfei,zhuoyazhang,guangcanliu}@seu.edu.cn
jyma2010@gmail.com

Abstract

Video fusion integrates complementary information in multiple source video streams into a single sequence for complete scene perception. Existing methods perform frame-wise fusion and lack unified spatio-temporal modeling across sequences, damaging temporal coherence across sequences. When dealing with degradations, they also ignore temporal dynamics and overload a single network with diverse degradations, leading to limited and imbalanced degradation suppression. To address these issues, we propose V2VFusion, the first unified video-to-video fusion framework that directly generates temporally coherent, high-quality fused videos from degraded multi-source inputs. By reformulating video fusion as a video-to-video conditional generation problem in latent diffusion space, our approach holistically models spatio-temporal dependencies across entire sequences and intrinsically ensures sequence-level consistency. On this basis, a text-guided degradation-conditioned ControlNet bridges high-level textual degradation descriptions with low-level visual priors, enabling interpretable and controllable guidance for structure-preserving fusion beyond purely data-driven conditioning. To handle heterogeneous and compound degradations, a state-modulated degradation-aware hierarchical mixture-of-experts module decouples expert selection into diffusion-state-aware filtering and degradation-aware routing. It prevents unreliable routing and fosters expert specialization to mitigate representation conflicts among diverse degradations. Experiments on infrared-visible, multi-exposure, and multi-focus video fusion tasks and complex degradations validate that V2VFusion surpasses existing methods in fusion quality and temporal stability. The code is available at https://github.com/ChenDoubleJ/V2VFusion.

1 Introduction

Real-world visual perception is inherently acquired in continuous video streams rather than isolated images [9]. Due to sensing limitations, video captured by a single modality/setting often suffers from incomplete scene perception. Thus, video fusion aims to integrate complementary information from multiple video sequences while maintaining temporal consistency, enabling robust and informative visual understanding. It plays a critical role in numerous downstream applications, *e.g.*, unmanned aerial surveillance [11], autonomous driving [24, 32], intelligent monitoring systems [17, 18].

Existing fusion research has largely been confined to static image fusion. It naturally lays the tractable and foundational groundwork for handling video streams. Image fusion achieves significant progress with methods based on auto-encoder [12, 13], CNN [34, 35, 3], Transformer [23, 26, 38],

*Co-corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

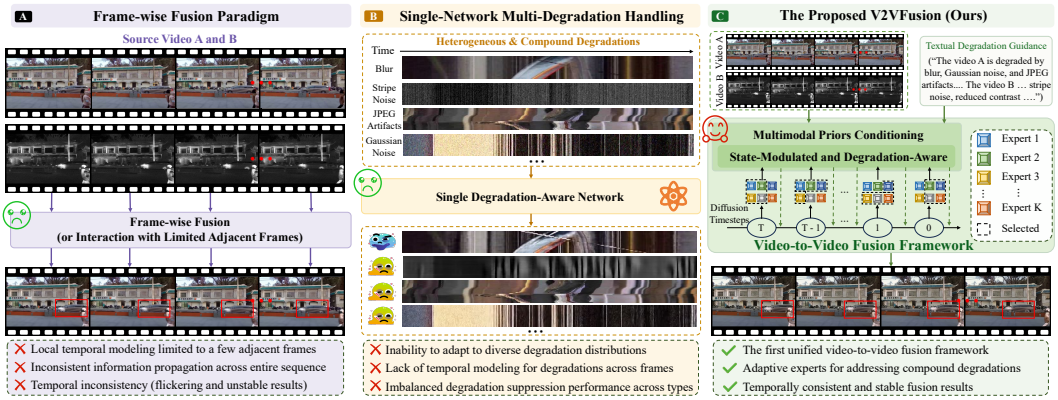

Figure 1: Existing video fusion paradigm. The fusion result in A is generated by VideoFusion [31].

GAN [21, 22], and diffusion models [42, 40, 14, 39]. While these approaches demonstrate impressive performance in single-frame image fusion, they lack explicit temporal modeling. When directly extended to video fusion, they often fail to maintain temporal continuity, leading to flickering artifacts and motion inconsistencies in fused videos, as shown in Fig. 1.

Recently, several video fusion methods have been proposed [47, 10, 31]. They follow a video-to-frame-to-video paradigm: source videos are first decomposed into short frame sequences. Rather than directly generating a temporally coherent fused video as a unified output, this paradigm generates temporally independent fused frames, which are reassembled into a fused video. Thus, temporal dependencies are only implicitly propagated through neighboring frames, resulting in short-range interactions and inconsistent information propagation across sequence. As a result, temporal inconsistencies accumulate over time, still producing flickering artifacts in fused videos. Although temporal consistency or optical flow-based constraints are introduced, they remain external corrections rather than a fundamental solution to the underlying frame-wise modeling paradigm. Thus, it remains challenging and pressing to design a unified video-level spatial–temporal modeling rather than frame-wise processing to produce temporally coherent and sequence-stable fusion results.

Moreover, due to imperfect imaging conditions or transmission processes, real-world video sequences are often affected by various degradations, such as noise, blur, and JPEG compression artifacts. Unlike static images, video degradations exhibit complex temporal behaviors. Some degradations persist across consecutive frames, while others fluctuate independently across frames, making degradation removal in video fusion more challenging. (1) While some degradation-aware image fusion methods [41, 43, 36, 37, 4] can deal with single-frame image degradation, they operate on individual frames independently. They fail to exploit the structural consistency across consecutive frames and ignore the temporal degradation evolution, leading to inconsistent suppression and temporal artifacts when directly applied to video with compound degradations. (2) While some data-driven video fusion methods [10, 31] can handle degraded inputs, their single network is unable to cope well in complex scenarios with compound degradations. Different degradations are heterogeneous in nature and affect contents differently. A single network cannot learn disentangled degradation-specific representations and effectively model heterogeneous degradations, making it struggle with complex degradation mixtures. Thus, designing a framework that models temporal structures while adaptively handling heterogeneous and compound degradations remains challenging.

To address these challenges, we propose V2VFusion, a novel text-controlled video-to-video diffusion framework for high-quality video fusion. Fundamentally departing from conventional frame-wise pipelines, we for the first time reformulate video fusion as a direct video-to-video conditional generation problem, enabling unified spatio-temporal modeling over entire video sequences and producing temporally coherent fused videos. Within this framework, we introduce a text-guided degradation-conditioned ControlNet that unifies multi-modal priors (i.e., degraded video and textual degradation description), bridging low-level visual cues with high-level language understanding to adaptively regulate restoration and fusion. This provides a flexible and interpretable control mechanism beyond purely data-driven conditioning. To further handle heterogeneous and compound degradations, we propose a state-modulated degradation-aware hierarchical mixture-of-experts module. Unlike

conventional MoE that routes experts solely based on input features, our design decomposes expert selection into a hierarchical process conditioned on both the diffusion state and the degradation type. It avoids unreliable expert activation in denoising stages and encourages experts to specialize in distinct degradation patterns, mitigating representation conflicts under degradation mixtures. Extensive experiments on multiple video fusion tasks under diverse degradations demonstrate that V2VFusion consistently outperforms existing methods. Our main contributions are summarized as follows:

• We propose the first video-to-video fusion framework that directly generates fused videos rather than relying on frame-wise fusion paradigms. By reformulating video fusion as a conditional video-to-video generation problem in the latent diffusion space, it holistically models spatio-temporal dependencies across entire sequences, intrinsically ensuring temporally coherent fusion results.

• We develop a text-guided degradation-conditioned ControlNet that bridges high-level semantic descriptions with low-level visual priors within the latent diffusion process. By unifying these multi-modal priors, it enhances degradation understanding and enables structure-preserving generation through interpretable and controllable guidance, moving beyond purely data-driven fusion paradigms.

• We propose a state-modulated degradation-aware hierarchical mixture-of-experts module to handle heterogeneous and compound degradations in video fusion. Instead of the conventional MoE that uniformly routes experts solely based on input features, it decouples expert selection into diffusion-state-aware filtering and degradation-aware routing. This hierarchical design prevents unreliable routing across different denoising stages and fosters expert specialization, thereby mitigating representation conflicts among heterogeneous degradations.

• We train a unified fusion model for multiple video fusion tasks, including infrared–visible, multi-exposure, and multi-focus fusion. Extensive experiments show its consistent performance across these tasks and robust fusion quality under diverse real-world degradations such as blur, Gaussian noise, stripe noise, and JPEG compression artifacts.

2 Related Work

Image Fusion. Early deep learning approaches typically employ auto-encoder [12, 13]. Subsequent studies have explored CNN [34, 35, 3, 5] and Transformer [23, 26] architectures, learning fusion strategies in a fully data-driven manner. Furthermore, by incorporating adversarial learning, generative adversarial networks (GANs) have also been introduced to image fusion [21, 22] while they still suffer from unstable training and potential mode collapse. Thus, diffusion-based models [42, 40, 14] have emerged as a promising alternative, demonstrating strong generative capability and more stable training behavior. However, these methods assume ideal input quality and overlook real-world degradations. While degradation-aware methods [29, 43, 36, 37, 4] address this, they are still image-level and ignore temporal video characteristics.

Video Fusion. Compared with image fusion, video fusion requires modeling temporal consistency across frames. Recent studies aim to address this challenge by introducing temporal alignment or optical flow estimation to exploit cross-frame information. For example, TemCoCo [10] enhances temporal consistency in video fusion by introducing visual–semantic collaboration and temporal cooperative learning. VideoFusion [31] builds a spatio-temporal collaborative network to jointly exploit cross-modal complementarity and temporal dependencies. UniVF [47] proposes a unified multi-frame learning framework that aggregates neighboring frames for general video fusion. However, they focus on cross-modal information aggregation and typically assume relatively high-quality input videos. When source videos suffer from complex degradations, their performance often degrades.

3 Proposed Method

3.1 Overall Architecture

The overall framework is shown in Fig. 2. Given degraded source videos $X_A, X_B \in \mathbb{R}^{F\times H\times W\times 3}$ (F, H, W denoting frame numbers, height, and width), the goal is to generate a high-quality fused video $\hat{Y} \in \mathbb{R}^{F\times H\times W\times 3}$ that preserves complementary information while suppressing degradations.

We first formulate video fusion as a conditional video-to-video generation problem by designing a latent diffusion-based video-to-video fusion framework. Built upon this framework, we introduce a

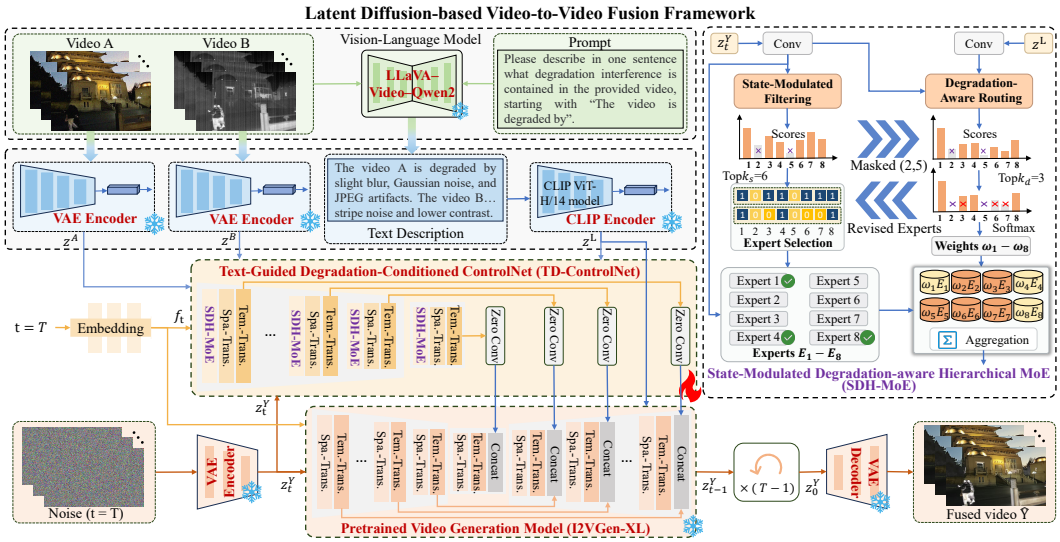

Figure 2: Overall framework of the proposed V2VFusion. The pretrained I2VGen-XL backbone adopts a 3D U-Net with interleaved spatial and temporal transformer blocks to jointly model spatial appearance and temporal dynamics. TD-ControlNet follows the ControlNet paradigm by duplicating the encoder and middle block of the pretrained backbone and injecting the extracted conditional features into the frozen denoising U-Net through zero-initialized convolutions.

text-guided degradation-conditioned ControlNet. It incorporates degraded video priors z^A, z^B and a textual degradation prior z^L into the diffusion process. By conditioning the generation on both visual cues and semantic guidance, the model can preserve structural information in source videos while explicitly reasoning about degradation characteristics, realizing degradation-robust video fusion.

Furthermore, to better handle compound degradations, we equip ControlNet with a state-modulated degradation-aware hierarchical mixture-of-experts module (SDH-MoE). Rather than relying on a single network to process diverse degradation patterns, this module adaptively routes intermediate features to specialized experts according to both diffusion states and degradation characteristics, enhancing the model's capability to disentangle and suppress multiple degradation types.

3.2 Latent Diffusion-based Video-to-Video Fusion Framework

Rather than relying on frame-by-frame generation, we aim to directly model video fusion as a unified video generation problem. Inspired by recent Stable Diffusion-derived video generation models [27, 45], we formulate video fusion as a video-to-video conditioned generation task within a latent diffusion framework, enabling holistic end-to-end video fusion.

First, we employ a pretrained VAE encoder $\mathcal{E}$ [1] to encode the ground-truth fused video Y into a latent representation $z^Y = \mathcal{E}(Y)$, where $z^Y \in \mathbb{R}^{F\times h\times w\times C}$. Then, Gaussian noise is progressively added to z^Y in forward process, modeled as a Markov chain with T timesteps. The forward process defines a conditional distribution $q(\cdot)$, representing the noise corruption process, formulated as:

$$q(z_t^Y|z_0^Y) = \mathcal{N}\left(z_t^Y; \sqrt{\hat{\alpha}_t} z_0^Y, \hat{\beta}_t \mathbf{I}\right), \tag{1}$$

where $\mathbf{I}$ is an identity matrix. $\mathcal{N}(0;\mathbf{I})$ is a standard normal distribution. β_t is the variance schedule at timestep t. $\hat{\beta}_t = \prod_{i=1}^{t} \beta_i$, $\hat{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$. z_t^Y is the latent noise at timestep t. As t increases, β_t gradually grows and $q(z_t^Y|z_0^Y)$ progressively approaches a standard normal distribution.

During the reverse denoising process, diffusion model starts from timestep T and progressively generates a high-quality fused video from Gaussian noise through a denoising network ϵ_θ. To enable controllable and degradation-aware fusion, we further introduce two conditional priors: *degraded video priors* and *textual degradation priors*. Degraded videos provide scene-level visual priors to preserve spatial structures and motion dynamics of source observations. Textual descriptions provide

high-level priors that explicitly indicate degradation types. By jointly leveraging complementary sources, the model can simultaneously reason about scene content and degradation characteristics.

Since priors originate from different modalities, representations are inherently heterogeneous. To enable effective conditioning within diffusion framework, both modalities are projected into latent space for feature alignment with diffusion variables. *For degraded video priors:* X_A, X_B are encoded by $\mathcal{E}$ to obtain latent representations z^A and z^B. *For textual degradation priors:* a multi-modal vision–language model, LLaVA-Video-Qwen2 [46] analyzes the degraded source videos and generates degradation descriptions (*e.g.*, *"The visible video is degraded by $\cdots$. The infrared video is degraded by $\cdots$."*) through an instructional prompt. Then, the textual descriptions are encoded by the pretrained CLIP text encoder (i.e., CLIP-ViT [28]) to obtain a latent representation z^L.

To incorporate these priors into pretrained video diffusion model, we design conditional guidance via ControlNet framework [44], which is primarily designed for image generation with single-modality control signals (e.g., edges or depth maps). When directly applied to video fusion, it faces two key limitations. First, it lacks mechanisms to jointly integrate heterogeneous conditional information from multiple modalities. Second, it does not explicitly model degradation characteristics, making it difficult to provide effective guidance under compound degradation conditions.

Thus, we develop a text-guided degradation-conditioned ControlNet (TD-ControlNet) π_θ. It jointly incorporates degraded video priors and textual degradation semantics into the diffusion process. By bridging low-level visual degradation cues with high-level semantic descriptions, TD-ControlNet enables the diffusion model to adaptively regulate restoration and fusion (TD-ControlNet is detailed in Sec. 3.3). To preserve the capability of the pretrained video generation model while enabling effective conditional control, the outputs of TD-ControlNet are connected to the backbone through zero-initialized convolution layers. Finally, the reverse denoising process can be formulated as:

$$p_\theta(z_{t-1}^Y|z_t^Y, z^A, z^B, z^L) = \mathcal{N}(z_{t-1}^Y; \mu_\theta(z_t^Y, t, z^L, \pi_\theta(z_t^Y, z^A, z^B, t, z^L))), \sigma_t^2\mathbf{I}), \tag{2}$$

where μ_θ is the mean and calculated by denoising network ϵ_θ. The variance σ_t^2 is a time-dependent constant. p_θ denotes a posterior distribution. Through Bayes' theorem, the true reverse posterior distribution can be derived as $p_\theta(z_{t-1}^Y|z_t^Y, z_0^Y)$. Thus, the mean $\hat{\mu}_t$ can be acquired in timestep t by:

$$\hat{\mu}_t(z_t^Y, z_0^Y, z^A, z^B, z^L) = \frac{\sqrt{\hat{\alpha}_{t-1}}\beta_t}{1-\hat{\alpha}_t} z_0^Y + \frac{\sqrt{\alpha_t}(1-\hat{\alpha}_{t-1})}{1-\hat{\alpha}_t} z_t^Y. \tag{3}$$

However, the clean latent variable z_0^Y is unknown during inference. Therefore, diffusion models employ a denoising network ϵ_θ to predict the noise component and estimate the clean sample:

$$\hat{z}_0^Y = \frac{z_t^Y - \sqrt{1-\hat{\alpha}_t}\epsilon_\theta(z_t^Y, t, z^L, \pi_\theta(z_t^Y, z^A, z^B, t, z^L)))}{\sqrt{\hat{\alpha}_t}}. \tag{4}$$

Substituting the estimated clean sample $\hat{z}_0^Y$ into the posterior mean formulation yields the practical parameterization used in diffusion models:

$$\mu_\theta(z_t^Y, t, z^A, z^B, z^L) = \frac{z_t^Y - \frac{1-\alpha_t}{\sqrt{1-\hat{\alpha}_t}}\epsilon_\theta(z_t^Y, t, z^L, \pi_\theta(z_t^Y, z^A, z^B, t, z^L)))}{\sqrt{\alpha_t}}, \quad \sigma_t^2 = \frac{1-\hat{\alpha}_{t-1}}{1-\alpha_t}\beta_t. \tag{5}$$

The above formulation enables the model to iteratively predict the reverse transition and progressively remove noise, yielding the final latent representation z_0^Y. The generated fused video $\hat{Y}$ is then obtained by decoding z_0^Y into the pixel space using the VAE decoder [1].

3.3 Text-Guided Degradation-Conditioned ControlNet

To enable controllable and degradation-aware video fusion, we introduce a text-guided degradation-conditioned ControlNet (TD-ControlNet) that explicitly modulates the pretrained video diffusion model using both visual degradation cues and high-level textual degradation semantics. The key idea is to bridge low-level corrupted observations with semantic understanding, allowing the generation process to be guided by interpretable degradation priors rather than purely data-driven embeddings. Despite its effectiveness, representing diverse degradation types within a single unified representation space remains inherently limited.

Mixture-of-Experts (MoE) [48] provides a natural mechanism for modeling heterogeneous patterns by dynamically activating experts to handle compound degradations. However, effective routing typically

relies on explicit conditional cues, while degradation information (e.g., blur or JPEG artifacts) is not always observable in latent diffusion representations. Moreover, the information structure of the latent variable z_t^Y varies across diffusion timesteps. In early stages (large t), z_t^Y is close to Gaussian noise and lacks discriminative structures, whereas in later stages (small t), textures gradually emerge, making degradations more identifiable. Consequently, applying a conventional MoE uniformly across the diffusion process may lead to unreliable routing due to insufficient degradation cues.

It motivates a more structured formulation of degradation modeling in diffusion models. Two factors should be jointly considered instead of the formulation as a unified task. First, the diffusion state determines recoverable information (e.g., noise or fine details) contained in the latent variable. Second, the degradation type determines the appropriate processing strategy. In essence, reliable degradation identification depends on an interpretable diffusion state, as the model can only infer degradation types once the currently observable information in latent space is determined. *From a probabilistic perspective*, expert selection can thus be decomposed into two components: a diffusion-state-dependent candidate expert set and degradation-aware routing within this set, formulated as:

$$P(E \mid z_t, f^c) = P(E \mid f^c, E \in \mathcal{S}(z_t))\, P(E \in \mathcal{S}(z_t) | z_t), \tag{6}$$

where z_t denotes the latent variable at timestep t, f^c represents the degradation-aware conditional representation, and $\mathcal{S}(z_t)$ denotes the candidate expert subset determined by the diffusion state.

State-Modulated Degradation-aware Hierarchical Mixture-of-Experts (SDH-MoE)
Motivated by the above formulation, this module first performs a diffusion state-aware filtering step to determine which experts are compatible with the current denoising stage, and then conducts degradation-aware routing to select the most appropriate experts for handling the observed compound degradations. Specifically, in the state-aware filtering stage, given the latent feature h_t derived from the diffusion state z_t by in-layers, compatibility scores are predicted for each expert:

$$h_t = \text{in_layers}(z_t), \quad s = g_s(h_t), \quad s = [s_1, s_2, \ldots, s_K] \in \mathbb{R}^K, \tag{7}$$

where $g_s(\cdot)$ denotes the state-aware filtering network and K is the total number of experts. Each score s_i measures the compatibility between the i-th expert and the current diffusion state. Based on these scores, a binary filtering mask $m \in \{0,1\}^K$ is constructed using a Top-k operator. When $i \in \text{TopK}(s, k_s)$, $m_i = 1$. Otherwise, $m_i = 0$, where k_s denotes the number of candidate experts retained by the state-aware filtering stage. The resulting candidate expert subset can therefore be written as $\mathcal{S}(z_t) = \{i \mid m_i = 1\}$. This diffusion-state-aware mask acts as a filtering mechanism that restricts the expert search space to those experts compatible with the current diffusion stage. As a result, unreliable expert activation can be avoided when the latent representation still contains insufficient structural information.

After obtaining the candidate expert subset $\mathcal{S}(z_t)$ through state-aware filtering, the degradation-aware router further performs fine-grained expert selection based on the degradation semantic representation f^c. Specifically, the router predicts degradation routing scores for all experts as:

$$f^c = \text{Context-Layers}(z^L), \quad r = g_d\big(h_t + f^c\big), \quad r = [r_1, r_2, \ldots, r_K] \in \mathbb{R}^K, \tag{8}$$

where $g_d(\cdot)$ is the degradation-aware gating network. Context-Layers (i.e., mean, SiLU and Linear) project the textual degradation representation into latent feature space, enabling alignment with h_t. To ensure that expert selection is restricted to the candidate subset determined by the diffusion state, the routing scores are filtered corresponding to $\mathcal{S}(z_t)$. When $i \in \mathcal{S}(z_t)$, $\tilde{r}_i = r_i$. Otherwise, $\tilde{r}_i = -\infty$. Based on it, the router performs a Top-k selection within the candidate expert subset:

$$\mathcal{T}(z_t, f^c) = \text{TopK}(\tilde{r}_i \mid i \in \mathcal{S}(z_t), k_d), \tag{9}$$

where k_d denotes the number of experts activated by the degradation-aware router. Subsequently, the routing weights are then computed using a softmax function over the selected experts:

$$\omega_i = \frac{\exp(\tilde{r}_i)}{\sum j \in \mathcal{T}(z_t, f^c) \exp(\tilde{r}_j)}, \quad i \in \mathcal{T}(z_t, f^c), \tag{10}$$

where $\omega_i = 0$ for $i \notin \mathcal{T}(z_t, f^c)$. Finally, the SDH-MoE output is obtained by aggregating the outputs of the selected experts weighted by the routing coefficients:

$$h_t^{\text{moe}} = \sum\nolimits_{i \in \mathcal{T}(z_t, f^c)} \omega_i E_i(h_t), \tag{11}$$

where $E_i(\cdot)$ denotes the i-th expert network. Through this hierarchical routing design, SDH-MoE dynamically activates a compact subset of experts that are jointly compatible with the diffusion state and specialized for the underlying degradation, thereby improving routing reliability.

3.4 Loss Functions

To accelerate the diffusion sampling process, we adopt the progressive distillation paradigm to reduce the sampling cost. The optimization objective is defined as:

$$\mathcal{L}_{\text{Diff}} = \mathbb{E}_{z_0^Y, \epsilon \sim \mathcal{N}(0,I), t} \left[\left\| \sqrt{\hat{\alpha}_t}\epsilon - \sqrt{1-\hat{\alpha}_t} z_0^Y - \epsilon_\theta \left(z_t^Y, t, z^A, z^B, z^L \right) \right\|_2^2 \right]. \tag{12}$$

To improve structural fidelity and suppress degradation artifacts, we introduce a frequency reconstruction loss. At timestep t, the generated fused video $\hat{Y}$ and ground-truth Y are transformed into the frequency domain by Discrete Fourier Transform. A low-pass filter η yields low/high frequency components as $\hat{f}_l = \mathcal{F}(\hat{Y}) \odot \eta$, $\hat{f}_h = \mathcal{F}(\hat{Y}) \odot (1-\eta)$, and similarly f_l, f_h for Y, where $\mathcal{F}(\cdot)$ is DFT. $\odot$ is Hadamard product. With a weighting control parameter ψ, the loss is defined as:

$$\mathcal{L}_{\text{Freq}} = (\frac{t}{T})^\psi \|f_l - \hat{f}_l\|_2^2 + (1 - (\frac{t}{T})^\psi) \|f_h - \hat{f}_h\|_2^2. \tag{13}$$

Finally, with λ being a balancing coefficient, the overall loss function is $\mathcal{L} = \mathcal{L}_{\text{Diff}} + \lambda \mathcal{L}_{\text{Freq}}$.

4 Experiment

Datasets and Implementation. We train and evaluate across three video fusion tasks: infrared–visible video fusion on M3SVD [31], multi-exposure video fusion on YouTube-HDR [47], and multi-focus video fusion on DAVIS [47]. Various degradations are introduced during training, including blur, Gaussian noise, stripe noise, and JPEG compression artifacts. Based on high-quality source videos and existing video fusion methods [31, 10, 47], we generate candidate fused videos and select high-quality results as ground truth. More details are provided in *Supplementary Material*.

We train *a unified fusion model* for multiple video fusion tasks. We use AdamW optimizer with initial learning rate of 5×10^{-5}, scheduled with 500 warmup steps followed by cosine annealing. Input videos are sampled at 30 FPS with a duration of 1 second. Hyperparameters are $\psi = 2$, $\lambda = 0.5$, $K = 8$, $k_s = 6$, $k_d = 3$, $T = 999$. The frequency mask η is determined by the 80th-percentile magnitude threshold. Experiments are conducted on four NVIDIA A800 GPUs with a batch size of 4 and trained for 15,000 iterations.

Results and Analysis. We compare V2VFusion with several SOTA image fusion methods (U2Fusion [34], DeFusion [16], TC-MoA [49], FusionBooster [8], PromptFusion [19], DCEvo [20], GIFNet [7], Mask-Difuser [30]) and video fusion approaches (VideoFusion [31], TemCoCo [10], and UniVF [47]). For fair comparison under degradation conditions, we employ several SOTA pre-enhancement methods, including visible denoising [6], infrared denoising [2], deblurring [25], and JPEG artifact removal [15], which are retrained under default experimental settings.

Qualitative Comparison. Results across multiple video fusion tasks and diverse degradation conditions are shown in Figs. 3-5. More qualitative comparisons and additional evaluations on challenging temporal scenarios, including fast motion and long video sequences, are provided in the *Supplementary Material*. We analyze the results from three aspects.

1) Fusion Quality: our method achieves clearer structures and richer textures. Existing methods often produce flickering and texture distortions due to limited temporal consistency, whereas our video-level generation framework achieves more stable coherence while preserving fine spatial details.

2) Degradation Suppression: our method is more effective at mitigating compound degradations. Although some existing fusion pipelines combine pre-enhancement modules with fusion networks, residual degradations still remain due to the limited synergy between enhancement and fusion. In contrast, our unified text-guided multi-degradation-aware framework jointly performs restoration and fusion, thereby suppressing various degradations more effectively.

3) Temporal Consistency: our method demonstrates superior temporal consistency. i) Temporal profiles: In Fig. 6, competitors exhibit noticeable temporal noise and structural discontinuities with severe fluctuations, whereas our method yields the most stable temporal structures with substantially reduced flickering and degradation artifacts. *ii) Pixel intensity curves:* In Fig. 7(b), we assess temporal smoothness via pixel intensity variations across frames. Our method maintains consistent intensity transitions, while competitors display irregular fluctuations indicative of inferior temporal stability.

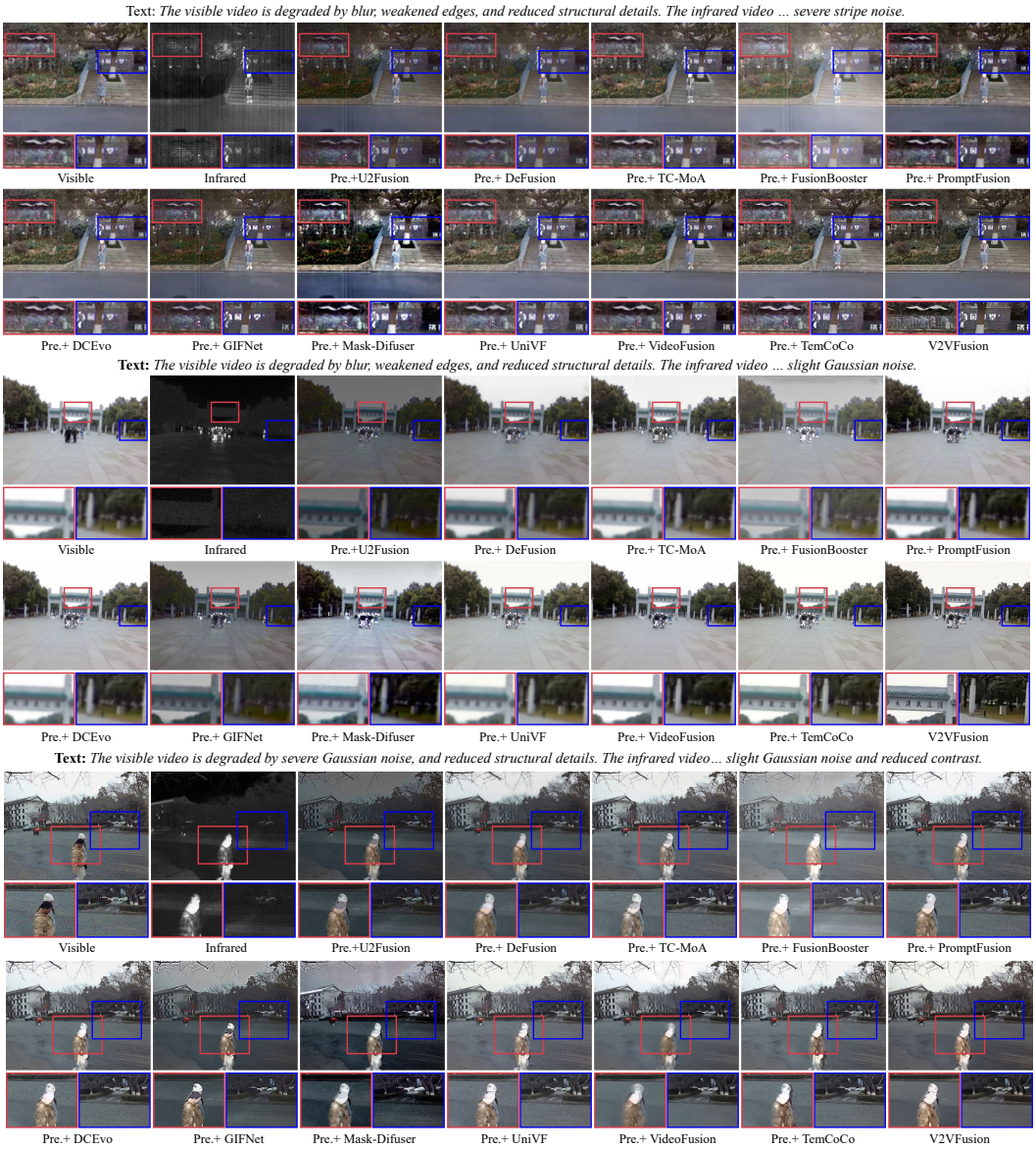

Figure 3: Results on M3SVD infrared–visible dataset (Pre.: pre-enhancement [6], [2], [25], [15]).

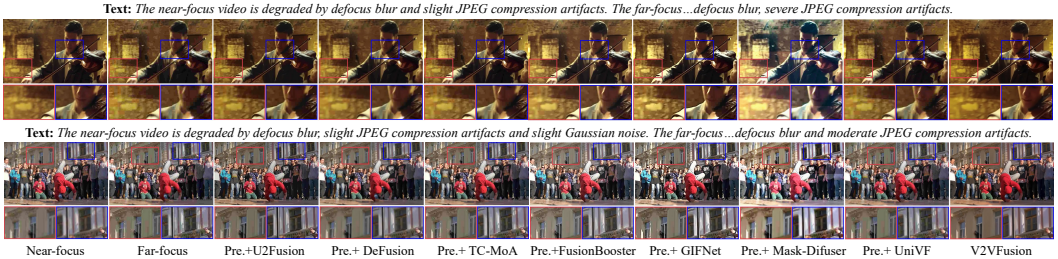

Figure 4: Results on DAVIS multi-focus dataset (Pre.: pre-enhancement [6, 2, 25, 15]).

iii) Inter-frame differences: In Fig. 7(c), we compute consecutive frame differences. Our method produces more motion-aligned difference patterns, whereas competitors exhibit motion artifacts and residual degradation, reflecting weaker temporal coherence.

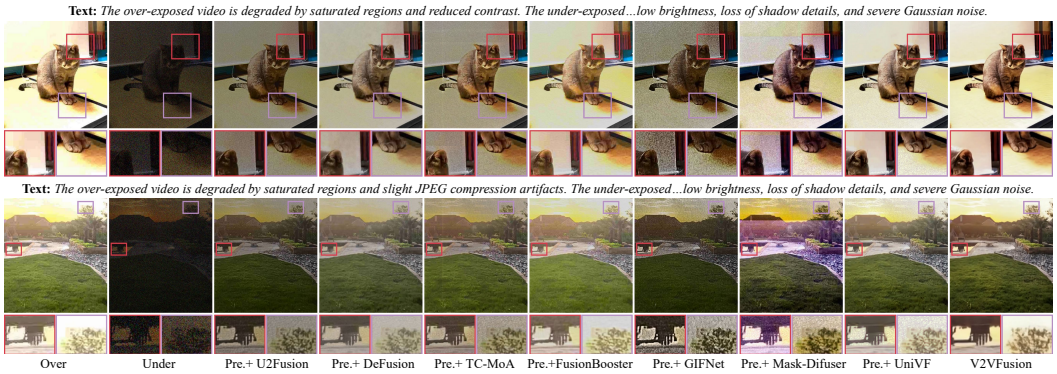

Figure 5: Results on YouTube. multi-exposure dataset (Pre.: pre-enhancement [6], [2], [25], [15]).

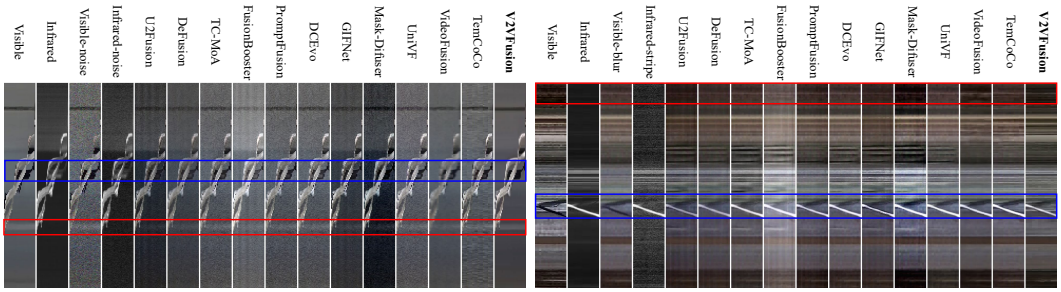

Figure 6: Comparison of temporal profiles (competing methods adopt pre-enhancement).

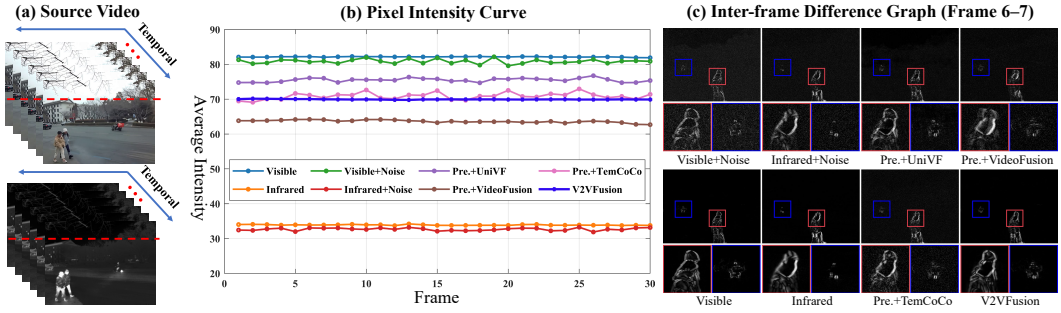

Figure 7: Comparison of pixel intensity curve and inter-frame difference (Pre.: pre-enhancement).

Table 1: Results on three video fusion tasks (P.: pre-enhancement; **red**: optimal; **blue**: suboptimal).

Metric	Infrared-Visible (M3SVD)											
	P.+U2Fus.	P.+DeFus.	P.+TC.	P.+FusionB.	P.+Prompt.	P.+DCEvo	P.+GIF.	P.+Mask.	P.+VideoFus.	P.+TemCoCo	P.+UniVF	**V2VFusion**
MI↑	1.982	2.224	2.002	2.118	2.210	2.457	2.433	2.190	2.523	**2.661**	2.31	**2.684**
Qabf↑	0.393	0.486	0.536	0.510	0.592	**0.671**	**0.653**	0.536	0.622	0.634	0.626	0.607
tSSIM↑	0.942	0.914	0.866	0.876	0.858	0.852	0.864	0.850	**0.972**	0.953	0.885	**0.983**
tF↓	1.402	1.496	2.285	2.288	2.268	2.331	2.502	2.830	**0.792**	1.404	2.256	**0.514**
tLP↓	0.352	0.535	0.536	0.578	0.599	0.580	0.610	0.614	**0.155**	0.354	0.493	**0.110**
BiSWE↓	3.502	3.155	4.456	4.608	4.559	4.660	5.000	5.591	**1.592**	3.503	4.025	**0.881**

Metric	Multi-Exposure (YouTube-HDR)								Multi-Focus (DAVIS)							
	P.+U2Fus.	P.+DeFus.	P.+TC.	P.+FusionB.	P.+GIF.	P.+Mask.	P.+UniVF	**V2VFusion**	P.+U2Fus.	P.+DeFus.	P.+TC.	P.+FusionB.	P.+GIF.	P.+Mask.	P.+UniVF	**V2VFusion**
MI↑	2.331	2.644	3.912	2.986	4.197	3.048	**4.318**	**4.495**	3.563	4.783	5.346	5.662	**6.343**	5.117	6.023	**6.489**
Qabf↑	0.369	0.577	0.544	0.613	**0.692**	0.435	0.686	**0.739**	0.615	0.553	0.634	**0.726**	0.712	0.520	0.701	**0.778**
tSSIM↑	0.592	0.683	0.542	0.681	**0.444**	0.495	0.626	**0.767**	0.598	**0.658**	0.585	0.553	0.542	0.575	0.577	**0.669**
tF↓	**4.157**	4.856	5.204	4.163	7.852	7.193	5.151	**3.882**	5.823	5.905	5.807	6.819	8.147	7.905	**4.871**	**4.710**
tLP↓	0.142	0.133	0.152	0.131	0.185	0.177	**0.127**	**0.075**	0.186	0.207	0.193	0.192	0.191	0.205	**0.177**	**0.116**
BiSWE↓	4.584	4.086	5.909	4.553	9.902	8.984	**4.578**	**2.566**	5.354	**4.595**	5.703	6.431	7.243	7.948	6.092	**3.231**

Quantitative Comparison. We evaluate with spatial quality metrics (MI, Qabf) to assess information fidelity and temporal consistency metrics (tSSIM, tF, tLP, BiSWE) to measure the stability of inter-frame representations and temporal fluctuations. As reported in Tab. 1, our method consistently

Table 2: Quantitative results of ablation study.

Index	Infrared-Visible (M3SVD)						Multi-Exposure (YouTube-HDR)						Multi-Focus (DAVIS)					
	MI↑	Qabf↑	tSSIM↑	tF↓	tLP↓	BiSWE↓	MI↑	Qabf↑	tSSIM↑	tF↓	tLP↓	BiSWE↓	MI↑	Qabf↑	tSSIM↑	tF↓	tLP↓	BiSWE↓
w/o $\mathcal{L}_{\text{Freq}}$	2.45	0.55	0.95	0.57	0.15	0.90	4.30	0.69	0.72	3.97	0.08	2.59	6.39	0.72	0.61	4.80	0.13	**3.22**
w/o Text	2.38	0.52	0.92	0.60	0.14	0.93	4.27	0.61	0.69	4.01	0.10	2.63	6.31	0.67	0.59	4.82	0.15	3.31
w/o SDH-MoE	2.53	0.56	0.97	0.58	0.12	0.89	4.38	0.67	0.70	3.90	0.11	2.58	6.29	0.68	0.59	4.78	0.16	3.27
w/o State	2.59	0.58	**0.99**	0.54	0.13	0.90	4.44	0.71	0.73	3.92	0.09	2.59	6.43	0.74	0.62	4.73	0.15	3.24
V2VFusion	**2.68**	**0.60**	0.98	**0.51**	**0.11**	**0.88**	**4.49**	**0.73**	**0.76**	**3.88**	**0.07**	**2.56**	**6.48**	**0.77**	**0.66**	**4.71**	**0.11**	3.23

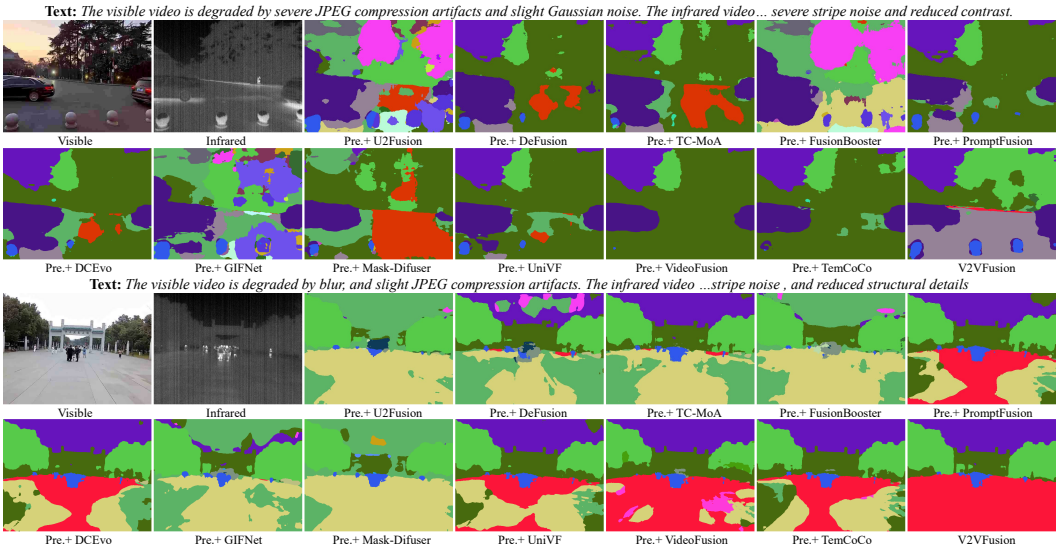

Figure 8: Qualitative comparison results of semantic segmentation (Pre.: pre-enhancement).

achieves superior performance across all benchmarks. Improvements in MI and Qabf indicate that the proposed framework can better preserve complementary information and structural details in source videos. Advantages in tSSIM, tF, tLP, and BiSWE show that our fused videos exhibit more stable temporal structures with fewer frame-to-frame inconsistencies and reduced flickering artifacts, validating the effectiveness of the proposed video-to-video generation paradigm (see *Supplementary Material* for the no-reference quality evaluation results).

Ablation Study. We conduct ablation studies by progressively removing different modules: **I**: removing the frequency reconstruction loss (w/o $\mathcal{L}_{\text{Freq}}$); **II**: removing the textual degradation description (w/o Text); **III**: removing the proposed SDH-MoE module (w/o SDH-MoE); **IV**: removing the state-modulated mechanism within SDH-MoE (w/o State). As shown in Tab. 2, removing individual components generally leads to a decline in overall performance across different tasks and evaluation metrics, demonstrating the contribution of each component to the effectiveness of the proposed framework (Qualitative comparisons are further provided in *Supplementary Material*).

Semantic Verification on High-Level Task. To further evaluate the effectiveness of our method in preserving and enhancing high-level semantic information, we conduct experiments on semantic segmentation using SegFormer [33]. As shown in Fig. 8, our method produces more semantically coherent and structurally consistent predictions, especially in regions with severe degradations. This indicates that the proposed fusion framework not only improves low-level visual quality but also effectively preserves discriminative features required for high-level semantic understanding.

5 Conclusion

This paper presents V2VFusion, a unified video-to-video fusion framework that generates temporally coherent results by conditional generation in latent diffusion space. A text-guided degradation-conditioned ControlNet and a state-modulated degradation-aware MoE respectively bridge semantic control and handle compound degradations. Experiments demonstrate its effectiveness.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62502085 and U21B2027, the Natural Science Foundation of Jiangsu Province of China under Grant BK20241280, and the Start-up Research Fund of Southeast University under Grant RF1028624061 and RF1028623006.

References

[1] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[2] L. Cai, X. Dong, K. Zhou, and X. Cao. Exploring video denoising in thermal infrared imaging: Physics-inspired noise generator, dataset, and model. *IEEE Transactions on Image Processing*, 33:3839–3854, 2024.

[3] B. Cao, Y. Sun, P. Zhu, and Q. Hu. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23555–23564, 2023.

[4] Z. Cao, Y. Zhong, Z. Wang, and L.-J. Deng. Mmaif: Multi-task and multi-degradation all-in-one for image fusion with language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11744–11754, 2025.

[5] J. Chen, H. Xu, G. Liu, and J. Ma. Breaking low-light fusion barrier: Unsupervised darkness and noise-aware visible and infrared image fusion network. *IEEE Transactions on Image Processing*, 35:7954–7969, 2026.

[6] Z. Chen, T. Jiang, X. Hu, W. Zhang, H. Li, and H. Wang. Spatiotemporal blind-spot network with calibrated flow alignment for self-supervised video denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2411–2419, 2025.

[7] C. Cheng, T. Xu, Z. Feng, X. Wu, Z. Tang, H. Li, Z. Zhang, S. Atito, M. Awais, and J. 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*, pages 28102–28112, 2025.

[8] C. Cheng, T. Xu, X.-J. Wu, H. Li, X. Li, and J. Kittler. Fusionbooster: A unified image fusion boosting paradigm. *International Journal of Computer Vision*, 133(5):3041–3058, 2025.

[9] J. Ding, H. Zhang, Z. Wang, J. Xiao, X. Tian, Z. Han, and J. Ma. Cmvf: Cross-modal unregistered video fusion via spatio-temporal consistency. *Information Fusion*, page 104212, 2026.

[10] M. Gong, H. Zhang, X. Yi, L. Tang, and J. Ma. Temcoco: Temporally consistent multi-modal video fusion with visual-semantic collaboration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14326–14335, 2025.

[11] S. Hu, W. Ni, X. Wang, and A. Jamalipour. Disguised tailing and video surveillance with solar-powered fixed-wing unmanned aerial vehicle. *IEEE Transactions on Vehicular Technology*, 71 (5):5507–5518, 2022.

[12] H. Li and X.-J. Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2018.

[13] H. Li, X.-J. Wu, and T. Durrani. Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Transactions on Instrumentation and Measurement*, 69(12):9645–9656, 2020.

[14] J. Li, C. Jiang, J. Jiang, P. Liang, J. Ma, and L. Nie. Towards unified semantic and controllable image fusion: A diffusion transformer approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[15] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.

[16] P. Liang, J. Jiang, X. Liu, and J. Ma. Fusion from decomposition: A self-supervised decomposition approach for image fusion. In *Proceedings of the European Conference on Computer Vision*, pages 719–735. Springer, 2022.

[17] H. Liu, S. Chen, and N. Kubota. Intelligent video systems and analytics: A survey. *IEEE Transactions on Industrial Informatics*, 9(3):1222–1233, 2013.

[18] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5802–5811, 2022.

[19] J. Liu, X. Li, Z. Wang, Z. Jiang, W. Zhong, W. Fan, and B. Xu. Promptfusion: Harmonized semantic prompt learning for infrared and visible image fusion. *IEEE/CAA Journal of Automatica Sinica*, 12(3):502–515, 2024.

[20] J. Liu, B. Zhang, Q. Mei, X. Li, Y. Zou, Z. Jiang, L. Ma, R. Liu, and X. Fan. Dcevo: Discriminative cross-dimensional evolutionary learning for infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2226–2235, 2025.

[21] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information fusion*, 48:11–26, 2019.

[22] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Transactions on Image Processing*, 29:4980–4995, 2020.

[23] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. 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.

[24] A. C. Muller and S. Narayanan. Cognitively-engineered multisensor image fusion for military applications. *Information Fusion*, 10(2):137–149, 2009.

[25] J. Pan, B. Xu, J. Dong, J. Ge, and J. Tang. Deep discriminative spatial and temporal network for efficient video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22191–22200, 2023.

[26] D. Rao, T. Xu, and X.-J. Wu. Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network. *IEEE Transactions on Image Processing*, 2023.

[27] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[28] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *In the Thirty-sixth Annual Conference on Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[29] L. Tang, Y. Deng, X. Yi, Q. Yan, Y. Yuan, and J. Ma. Drmf: Degradation-robust multi-modal image fusion via composable diffusion prior. In *Proceedings of the ACM International Conference on Multimedia*, pages 8546–8555, 2024.

[30] L. Tang, C. Li, and J. Ma. Mask-difuser: A masked diffusion model for unified unsupervised image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[31] L. Tang, Y. Wang, M. Gong, Z. Li, Y. Deng, X. Yi, C. Li, H. Xu, H. Zhang, and J. Ma. Videofusion: A spatio-temporal collaborative network for multi-modal video fusion and restoration. *arXiv preprint arXiv:2503.23359*, 2025.

[32] Y. Wen, Y. Zhao, Y. Liu, F. Jia, Y. Wang, C. Luo, C. Zhang, T. Wang, X. Sun, and X. Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6902–6912, 2024.

[33] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[34] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling. U2fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):502–518, 2020.

[35] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo. Fusiondn: A unified densely connected network for image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12484–12491, 2020.

[36] H. Xu, Y. Li, Y. Deng, J. Ma, and G. Liu. Deno-if: Unsupervised noisy visible and infrared image fusion method. In *the Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[37] H. Xu, X. Yi, C. Lu, G. Liu, and J. Ma. Urfusion: Unsupervised unified degradation-robust image fusion network. *IEEE Transactions on Image Processing*, 2025.

[38] H. Xu, J. Chen, X. Yi, P. Li, and J. Ma. Iterrf: Iterative and interpretable mutual enhancement for joint multi-modal image registration and fusion. *International Journal of Computer Vision*, 134(10):442, 2026.

[39] H. Xu, Y. Huang, L. Tang, J. Ma, and G. Liu. Diff-mef: Cross-modal diffusion framework with text prompts and semantic perception for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 35:3186–3201, 2026.

[40] X. Yi, L. Tang, H. Zhang, H. Xu, and J. Ma. Diff-if: Multi-modality image fusion via diffusion model with fusion knowledge prior. *Information Fusion*, 110:102450, 2024.

[41] X. Yi, H. Xu, H. Zhang, L. Tang, and J. 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*, pages 27026–27035, 2024.

[42] J. Yue, L. Fang, S. Xia, Y. Deng, and J. Ma. Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models. *IEEE Transactions on Image Processing*, 32: 5705–5720, 2023.

[43] H. Zhang, L. Cao, and J. Ma. Text-difuse: An interactive multi-modal image fusion framework based on text-modulated diffusion model. *In the Thirty-eighth Annual Conference on Advances in Neural Information Processing Systems*, 37:39552–39572, 2024.

[44] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[45] S. Zhang, J. Wang, Y. Zhang, K. Zhao, H. Yuan, Z. Qin, X. Wang, D. Zhao, and J. Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

[46] Y. Zhang, J. Wu, W. Li, B. Li, Z. Ma, Z. Liu, and C. Li. Video instruction tuning with synthetic data, 2024. *URL https://arxiv. org/abs/2410.02713*, 17.

[47] Z. Zhao, H. Bai, B. Ke, Y. Cui, L. Deng, Y. Zhang, K. Zhang, and K. Schindler. A unified solution to video fusion: From multi-frame learning to benchmarking. *arXiv preprint arXiv:2505.19858*, 2025.

[48] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Zhao, A. M. Dai, Q. V. Le, J. Laudon, et al. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

[49] P. Zhu, Y. Sun, B. Cao, and Q. Hu. Task-customized mixture of adapters for general image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7099–7108, 2024.

Technical Appendices and Supplementary Material

A Network Details of Pretrained Video Diffusion Model and TD-ControlNet

The pretrained video diffusion backbone adopts a 3D U-Net architecture composed of alternating spatial and temporal transformer blocks (Spa.-Trans. and Tem.-Trans.). Concretely, each spatial transformer module is followed by a temporal module, forming an interleaved structure that jointly models spatial appearance and temporal dynamics. The spatial transformer modules are inherited from Stable Diffusion 2.1, consisting of residual blocks, self-attention, and cross-attention layers for 2D feature extraction. Temporal modeling is achieved by dedicated temporal convolution and temporal attention layers inserted between spatial blocks. The temporal convolution is implemented as a 1D convolution with kernel size 3 along the temporal dimension, while the temporal attention is formulated as a 1D self-attention mechanism operating across frames. To preserve the pretrained spatial generative capability, all temporal transformer modules are initialized with zero output layers and then fine-tuned on video data. Besides, a U-Net skip-connection strategy is adopted in the 3D architecture, where temporally aggregated features from the encoder are directly passed to the corresponding decoder stages. These features are fused with decoder activations via concatenation, enabling effective reuse of hierarchical spatio-temporal representations and reconstruction consistency.

For TD-ControlNet, we follow the ControlNet paradigm by duplicating the encoder and middle block of the pretrained 3D U-Net, including architectures and parameters. The conditional inputs (i.e., degraded video representations and textual degradation embeddings) are fed into this copied branch to extract conditional features. These features are then injected into the frozen 3D U-Net through additional zero-initialized convolution layers, ensuring a smooth transition from unconditional generation to conditional control. Specifically, the output of the middle block in TD-ControlNet is added to the corresponding hidden representation in the middle block of original 3D U-Net. Similarly, the encoder features from the control branch are merged with the skip-connected features produced by the encoder of the backbone network. It follows the standard ControlNet formulation, while being further extended with the proposed state-modulated degradation-aware hierarchical MoE (SDH-MoE) module. Specifically, each expert in the MoE is implemented as a convolution-based network, enabling effective modeling of local spatio-temporal patterns. The MoE adaptively refines the injected conditional signals according to both the diffusion state and degradation characteristics, thereby improving the flexibility and robustness of addressing compound degradation conditions.

B Degradation Parameters

To improve the diversity of training samples and simulate realistic degradations in source videos, we randomly introduce different degradation types during data preparation.

Gaussian Noise. Gaussian noise is added with a standard deviation randomly sampled from $[40/255, 120/255]$. For each video clip, a noise level is first randomly determined and shared across the clip, while independent Gaussian noise is sampled and added to individual frames. This strategy preserves a consistent noise intensity across the video while maintaining frame-level stochastic variations, thereby better simulating realistic sensor noise.

Stripe Noise. It mainly occurs in infrared imaging systems. As the infrared videos in the M3SVD dataset already contain realistic stripe-noise degradation, we retain the original stripe noise provided by the dataset. Thus, no additional synthetic stripe noise is added to the M3SVD infrared videos.

Gaussian Blur. Gaussian blur is generated using kernel sizes randomly selected from $\{7, 9, 11, 13, 15, 17\}$, with the standard deviation σ randomly sampled from the range $[0.5, 3.0]$. The sampled blur parameters are applied to the corresponding source video clip. Since the visible videos in the M3SVD dataset already contain realistic blur degradation, we directly retain the original blur characteristics of these videos without introducing additional synthetic blur.

JPEG Compression Artifacts. JPEG compression artifacts are generated by JPEG encoding and decoding, with the quality factor randomly sampled from the range $[10, 50]$. To account for different temporal characteristics of compression degradation, the quality factor can either remain identical across all frames within a video clip or vary across individual frames. The former simulates temporally consistent compression artifacts, whereas the latter introduces temporally varying compression distortions that may occur in practical video acquisition and transmission scenarios.

C Pseudo Ground-Truth Construction

As high-quality ground-truth fused videos are unavailable for real-world video fusion, we construct pseudo ground-truth (pseudo GT) through a multi-stage **human–metric–VLM** collaborative curation procedure. It combines coarse human inspection, quantitative metric-based ranking, and vision-language-model (VLM) verification to reduce subjective bias and improve the reliability and consistency of pseudo GT.

i) Candidate generation. We first collect 6,000 aligned source-video pairs from 79 scenes, covering infrared-visible, multi-exposure, and multi-focus fusion tasks with diverse scene contents and motion patterns. For each source-video pair, the corresponding high-quality, undegraded videos are processed by three representative video fusion methods, i.e., TemCoCo, VideoFusion, and UniVF, to generate three candidate fused videos. Using multiple fusion methods helps reduce the bias introduced by any single method and provides diverse fusion priors for subsequent selection.

ii) Human-based coarse filtering. We then perform coarse manual inspection to remove clearly defective samples. Specifically, candidates with severe structural distortion, ghosting artifacts, obvious temporal flickering, or motion discontinuity are discarded. Manual inspection is only used for coarse quality filtering rather than directly determining the final pseudo GT. After this stage, 5,600 source-video pairs and their corresponding candidate results are retained.

iii) Metric-based quantitative selection. For each retained source-video pair, the candidate fused videos are further evaluated using both spatial and temporal fusion metrics, including MI, Qabf, tF, tSSIM, and BiSWE. Since these metrics have different numerical ranges and optimization directions, their scores are first normalized and then aggregated into an overall quality score. The candidate with the highest overall score is selected as the provisional pseudo GT. In this way, fine-grained candidate selection is determined by a unified quantitative criterion rather than subjective human preference.

iv) VLM-assisted quality verification. To further identify subtle perceptual and semantic defects that may not be fully captured by handcrafted metrics, we employ the pretrained LLaVA-Video-Qwen2 model for additional verification. A fixed prompt is used to assess complementary-information preservation, structural fidelity, temporal consistency, motion continuity, and residual fusion artifacts. The VLM assigns each provisional pseudo GT a quality score from 1 to 5 together with a brief justification, and samples with scores below 2 are discarded.

After the complete human–metric–VLM curation procedure, 5,119 source–fusion video pairs are retained as the final training data. By combining human perception, quantitative evaluation, and VLM-based semantic verification, the proposed construction procedure reduces dependence on subjective judgment and any single evaluation metric, while providing more reliable and diverse fusion supervision for model training.

D More Qualitative Comparison Results

We provide additional qualitative comparison results under three video fusion scenarios. More qualitative results of multi-focus video fusion are shown in Fig. 9. More qualitative results of infrared-visible video fusion are shown in Fig. 10. More qualitative results of multi-exposure video fusion are shown in Fig. 11. Overall, our method achieves superior degradation suppression and produces fused videos with clearer structures, richer textures, and more stable visual quality than existing approaches.

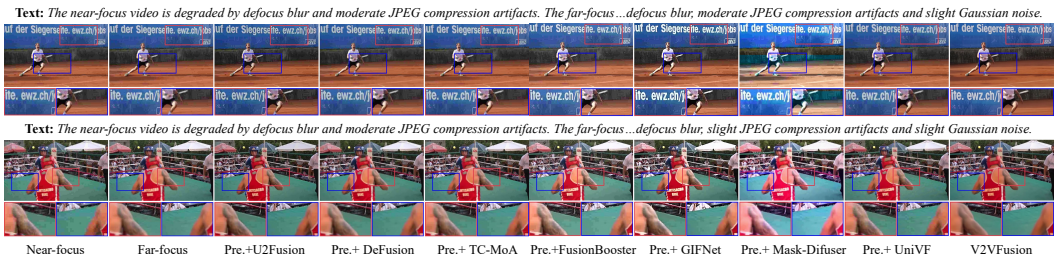

Figure 9: Results of multi-focus video fusion (Pre.: pre-enhancement [6], [2], [25], [15]).



Figure 10: Results of infrared-visible video fusion (Pre.: pre-enhancement [6], [2], [25], [15]).

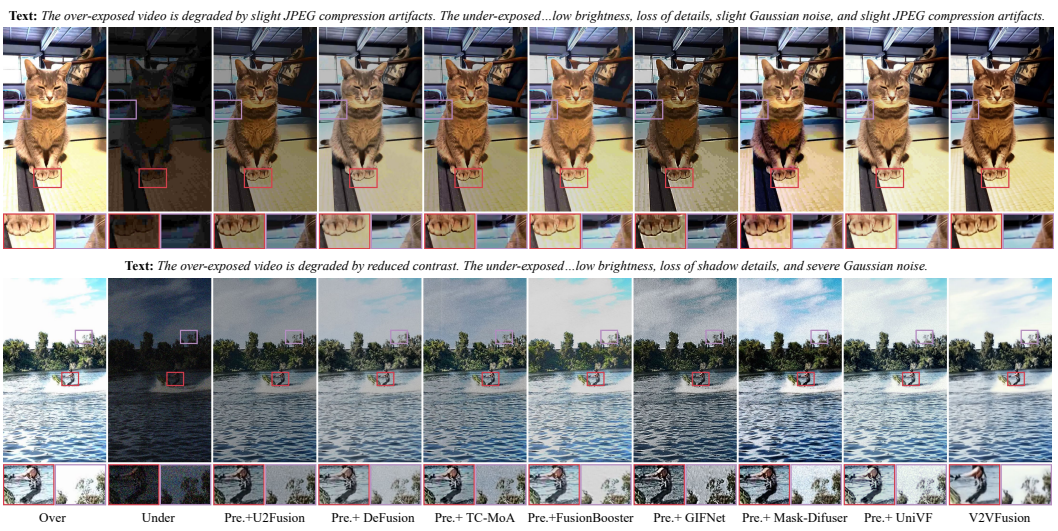

Figure 11: Results of multi-exposure video fusion (Pre.: pre-enhancement [6], [2], [25], [15]).

E Quantitative Results on No-Reference Quality Metrics

We further evaluate our method using four widely adopted no-reference image quality assessment metrics, including ILNIQE, NIQE, LIQE, and CLIP-IQA. Specifically, NIQE assesses image naturalness by measuring deviations from natural scene statistics, while ILNIQE employs richer statistical features to characterize diverse visual distortions. LIQE evaluates perceptual quality by jointly modeling image content, scene semantics, and distortion characteristics through vision–language representations. CLIP-IQA leverages the semantic priors of CLIP together with quality-related textual prompts to assess overall perceptual quality.

As shown in Tab. 3, our method consistently achieves the best overall performance across all three fusion tasks on the four no-reference quality metrics. These results further demonstrate that our method can generate fused videos with superior perceptual quality and more natural visual appearance.

Table 3: No-reference assessment on three video fusion tasks (P.: pre-enhancement; **red**: optimal).

Metric	Infrared-Visible (M3SVD)											
	P.+U2Fus.	P.+DeFus.	P.+DCEvo	P.+FusionB.	P.+GIFNet	P.+MaskD.	P.+PromptF.	P.+TCMoA	P.+VideoFus.	P.+TemCoCo	P.+UniVF	**V2VFusion**
ILNIQE↓	30.85	31.38	30.52	34.23	32.42	30.36	28.87	30.77	28.30	30.51	30.15	**24.02**
LIQE↑	1.38	1.46	1.90	1.41	1.44	1.74	1.86	1.81	1.57	1.92	1.69	**2.31**
CLIPIQA↑	0.37	0.33	0.38	0.35	0.38	0.36	0.37	0.38	0.38	0.40	0.35	**0.41**
NIQE↓	5.05	4.95	4.46	4.79	5.07	4.32	4.58	4.48	4.73	5.45	4.65	**4.20**

Metric	Multi-Exposure (YouTube-HDR)								Multi-Focus (DAVIS)							
	P.+U2Fus.	P.+DeFus.	P.+FusionB.	P.+GIFNet	P.+MaskD.	P.+TC-MoA	P.+UniVF	**V2VFusion**	P.+U2Fus.	P.+DeFus.	P.+FusionB.	P.+GIFNet	P.+MaskD.	P.+TC-MoA	P.+UniVF	**V2VFusion**
ILNIQE↓	29.73	**26.53**	29.53	32.80	26.63	26.90	27.06	**26.53**	27.46	29.69	25.13	**24.68**	24.69	24.69	25.37	**24.68**
LIQE↑	2.12	2.53	2.27	2.39	2.33	2.36	2.46	**2.56**	2.19	1.69	2.18	2.13	1.51	2.08	2.11	**2.21**
CLIPIQA↑	0.50	0.50	0.45	0.50	0.45	0.49	0.50	**0.51**	0.47	0.44	0.46	0.43	0.37	0.45	0.45	**0.50**
NIQE↓	4.68	3.72	4.25	4.82	3.65	3.65	3.74	**3.64**	4.68	4.67	4.93	4.61	4.30	4.31	4.37	**4.29**

F Qualitative Results of Ablation Study

We provide qualitative comparisons of the ablation study in Fig. 12. Specifically, we analyze several variants by removing key components, including the frequency reconstruction loss (w/o $\mathcal{L}_{Freq}$), textual degradation descriptions (w/o Text), the proposed SDH-MoE module (w/o SDH-MoE), and the state-modulated mechanism in SDH-MoE (w/o State). The results show that removing any of these components leads to noticeable degradation in both degradation suppression and fusion quality, further verifying the effectiveness of each component in our method.

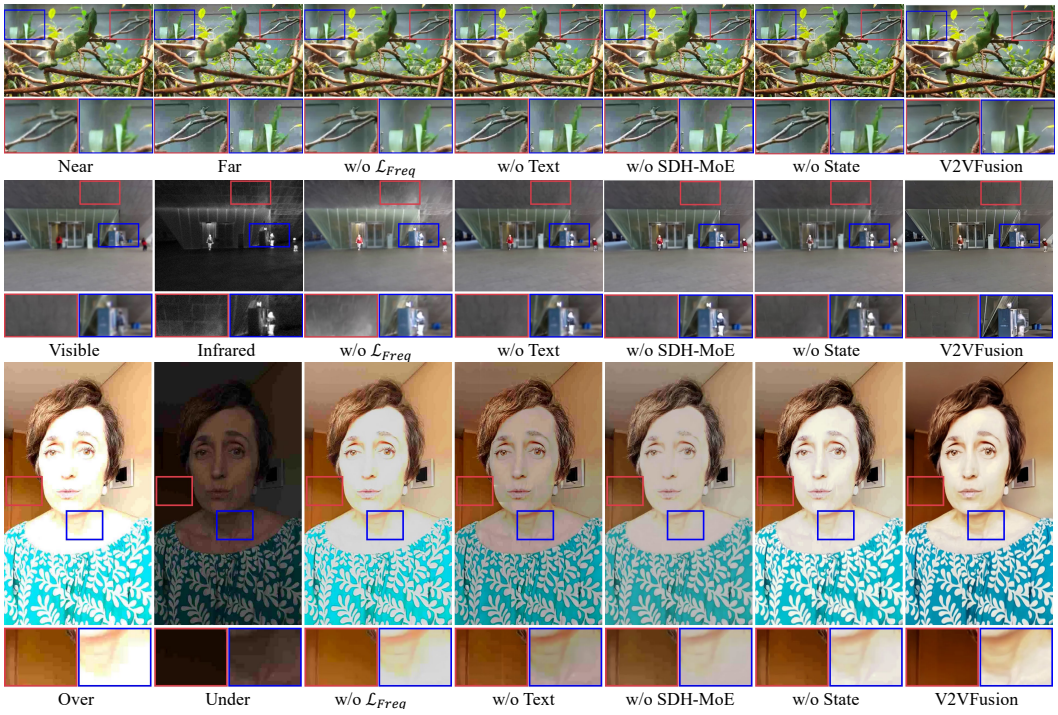

Figure 12: Qualitative result comparisons of the ablation study.

G Fast-Motion Fusion Evaluation

Video fusion requires not only preserving informative spatial details from different modalities, but also maintaining temporally consistent motion of dynamic objects across consecutive frames. While static regions should retain fine structural and texture details, moving targets should exhibit

coherent appearance and motion trajectories without introducing ghosting, misalignment, or temporal artifacts. These requirements are more challenging in fast-motion scenarios, where large inter-frame displacement and rapid temporal variation significantly increase the difficulty of temporal modeling.

To evaluate the robustness under such challenging conditions, we conduct qualitative comparisons on sequences containing fast-moving objects. As shown in Fig. 13, existing video fusion methods tend to suffer from noticeable structural distortions and motion-related artifacts in rapidly moving regions. In comparison, our method can preserve salient targets in the infrared modality while retaining detailed appearance information in the visible modality. Meanwhile, it maintains more stable target structures across consecutive frames, demonstrating stronger robustness to fast motion in video fusion.

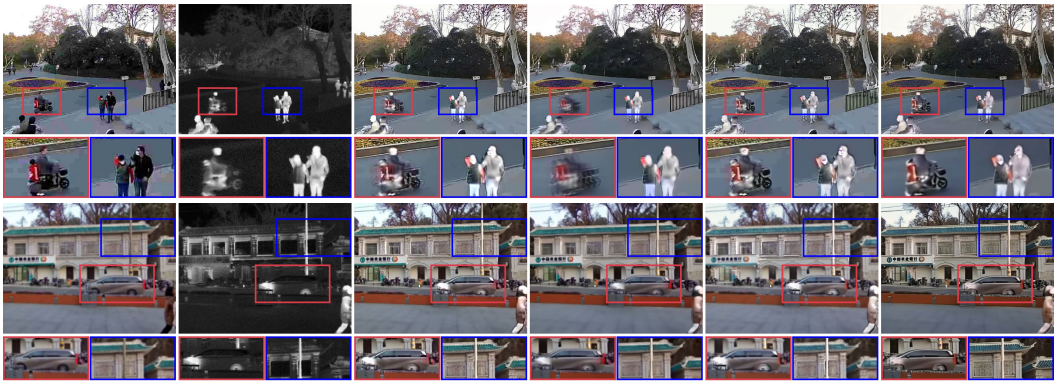

Figure 13: Qualitative comparison of visible-infrared video fusion under fast-motion scenarios.

H Long-Video Fusion Evaluation

To further evaluate the temporal robustness of different video fusion methods on long video sequences, we conduct experiments with video lengths of 90, 120, and 150 frames. We report four temporal consistency metrics, including tF, tLP, tSSIM, and BiSWE. As shown in Tab. 4, the performance of competing methods generally degrades as the video length increases, indicating that maintaining temporal consistency becomes increasingly challenging for long-term video fusion. In contrast, our method exhibits substantially better robustness across different sequence lengths.

Table 4: Comparisons on long videos with different numbers of frames.

Method	**90 Frames**				**120 Frames**				**150 Frames**			
	tF↓	**tLP↓**	**tSSIM↑**	**BiSWE↓**	**tF↓**	**tLP↓**	**tSSIM↑**	**BiSWE↓**	**tF↓**	**tLP↓**	**tSSIM↑**	**BiSWE↓**
TemCoCo	0.93	0.28	0.97	2.23	0.95	0.31	0.95	2.35	1.32	0.36	0.93	3.02
VideoFusion	0.64	0.23	0.97	1.13	0.73	0.29	0.94	1.35	1.02	0.32	0.92	2.79
UniVF	1.40	0.27	0.96	2.52	1.56	0.30	0.95	2.61	1.72	0.35	0.92	2.67
V2VFusion	**0.49**	**0.14**	**0.99**	**0.85**	**0.51**	**0.16**	**0.99**	**0.87**	**0.53**	**0.18**	**0.98**	**0.90**

I Limitations

Although the proposed framework demonstrates strong performance on multiple degradation-conditional fusion scenarios, the range of degradations considered in this work is primarily focused on common types (e.g., blur, Gaussian noise, stripe noise, and JPEG artifacts). Degradations that may arise from specific sensor characteristics or environmental conditions during imaging are not yet included. This limitation mainly stems from the scarcity of large-scale real-world multi-source video fusion datasets containing diverse degradation types. In future work, we plan to address this limitation by constructing a more comprehensive multi-source video fusion dataset with diverse real-world degradations. Such datasets would facilitate more reliable evaluation and promote the development of more robust video fusion methods under complex real-world conditions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction of the paper clearly state the claims made, including the contributions made in the paper and important assumptions and limitations.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations of the work performed by the authors, providing a balanced view of the study's scope and potential areas for improvement.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All the theorems, formulas, and proofs in the paper have been numbered and cross-referenced, and all assumptions have been clearly stated or referenced in the statement of any theorems.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper has fully disclosed all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and conclusions of the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will open access to the code after the paper is accepted.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have specified all the training and test details necessary to understand the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars and other appropriate information about the statistical significance of the experiments, ensuring the reliability and validity of the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computer resource for experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms in every respect with the NeurIPS Code of Ethics, ensuring ethical standards are upheld throughout the study.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both potential positive societal impacts and negative societal impacts of the work performed, providing a comprehensive evaluation of the study's broader implications.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The paper describes safeguards that have been put in place for the responsible release of data or models that have a high risk for misuse, ensuring that appropriate measures are taken to mitigate potential risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets used in the paper are properly credited, and the license and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer:[Yes]

    Justification: The new assets introduced in the paper are well documented, and the documentation is provided alongside the assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in the paper does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.