UAV-VIVID: A LARGE-SCALE UAV INFRARED–VISIBLE VIDEO DATASET AND A TEXT-GUIDED FRAME-LEVEL FUSION BASELINE

Anonymous authors
Paper under double-blind review

ABSTRACT

Infrared–visible (IR–VI) fusion is commonly developed on independent, well-aligned image pairs, leaving its behavior on continuous UAV streams insufficiently characterized. We introduce UAV-VIVID, a UAV-acquired IR–VI video dataset containing 375 temporally continuous scene segments and 190,978 paired frames. Its 96,087 registration-processed lower-altitude pairs define the fusion track evaluated in this work, while 79,992 higher-altitude pairs and 14,899 unregistered lower-altitude pairs are released as auxiliary resources. The dataset preserves temporal order, adopts scene-disjoint splits, and provides structured textual descriptions. To establish a controlled frame-level reference for this video setting, we further present the Residual Text-Guided Fusion Network (RT-FuseNet), which processes each IR–VI pair independently without neighboring frames or temporal state. Adaptive Residual Balancing regulates cross-modal interactions, while Text-Affine and Gated Modulation injects language guidance through channel-wise affine parameters and spatial gates. Experiments across multiple datasets evaluate frame-wise fusion quality, sequence-level output behavior, text conditioning, downstream utility, and the clean-input regularization effect of training-time dual-modal PGD augmentation.

INTRODUCTION

Infrared–visible (IR–VI) fusion combines the complementary sensing properties of infrared and visible imagery into a single informative representation. Infrared imagery captures thermal responses and remains effective under weak illumination, whereas visible imagery provides richer structures, textures, and color information under favorable lighting conditions. Their combination benefits downstream tasks such as object detection Liu et al. (2022) and semantic segmentation Liu et al. (2023a). Nevertheless, most existing fusion methods are developed and evaluated using independent image pairs Wang et al. (2025a); Li et al. (2025b); Chen et al. (2025), while existing protocols provide limited support for systematically examining their frame-by-frame output behavior on continuous sequences.

Practical UAV systems acquire infrared and visible observations as continuous paired streams, where neighboring frames exhibit strong correlations in scene content, target appearance, illumination, and imaging conditions Qiu et al. (2026); Tang et al. (2026). Existing IR–VI video benchmarks have begun to preserve such temporal structure, yet they differ substantially in acquisition platform, altitude coverage, and alignment resources. As illustrated in Fig. 1, representative benchmarks such as M3SVD, VF-Bench, and EUM3D each cover only part of this design space, whereas UAV-VIVID combines native UAV acquisition, temporally continuous video, explicitly separated altitude regimes, both registration-processed and raw cross-modal pairs, and structured textual descriptions

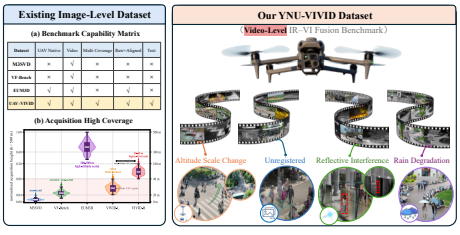


Figure 1: Motivation and representative UAV acquisition challenges in UAV-VIVID, including scale variation, residual misalignment, reflections, and rain.

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.

within a unified benchmark. In particular, the low- and high-altitude subsets expose a measurable acquisition shift associated with changing target scale and scene coverage, while the raw sequences further preserve registration uncertainty encountered in practical UAV sensing. UAV-VIVID additionally contains challenging conditions such as reflective interference and rain degradation. Rather than treating these observations as unordered image pairs, the benchmark preserves their scene and sequence organization through scene-disjoint partitions, enabling systematic evaluation of frame-wise fusion quality, cross-altitude behavior, registration robustness, and within-scene temporal consistency. This organization provides a common testbed for comparing frame-level and sequence-aware fusion methods under realistic UAV deployment variations.

MSRS Tang et al. (2022), LLVIP Jia et al. (2021), M3FD Liu et al. (2022), and FMB Liu et al. (2023a) provide established benchmarks for image fusion and downstream perception. Their standard protocols, however, primarily organize aligned image pairs rather than continuous, scene-grouped sequences. Consequently, they provide limited support for scene-disjoint evaluation, within-scene analysis of output variation, and systematic study of frame-level methods on temporally ordered UAV data.

Accordingly, we introduce UAV-VIVID, a large-scale UAV-acquired IR–VI video dataset collected using a DJI Matrice 4T. All drone flights and data acquisition operations were conducted by personnel certified by the Civil Aviation Administration of China. UAV-VIVID contains 375 continuous scenes and 190,978 temporally paired frames, organized into registration-processed lower-altitude fusion sequences, an unregistered lower-altitude daytime challenge subset, and higher-altitude auxiliary sequences. Scene-disjoint partitions keep all frames from the same continuous scene within a single split, thereby preventing correlated neighboring frames from crossing the training–test boundary.

The dataset covers daytime and nighttime acquisition, campus and urban environments, rainy conditions, reflective interference, and scenes containing small or partially occluded targets. Registration-processed lower-altitude sequences provide relatively large targets and detailed structures for pixel-level fusion evaluation, whereas the unregistered subset supports future research on registration uncertainty and cross-modal misalignment. Higher-altitude sequences provide wider aerial views and smaller targets as an auxiliary resource for tracking and other aerial perception tasks. All released sequences are further accompanied by structured frame-level descriptions containing global scene captions, object-aware descriptions, and region-level semantic cues, supporting language-guided fusion and multimodal perception research.

We also present the Residual Text-Guided Fusion Network (RT-FuseNet), a temporally agnostic frame-level reference evaluated under the proposed sequence-oriented protocol. Adaptive Residual Balancing (ARB) suppresses unreliable cross-modal interactions, whereas Text-Affine and Gated Modulation (TAGM) maps vision–language embeddings to channel-wise affine parameters and spatial gates. A source-driven objective preserves source intensity, edges, and structural information, while bounded dual-modal PGD is used only as a training augmentation. RT-FuseNet processes each ordered frame pair independently and therefore does not explicitly model temporal dependencies. Instead, it provides a controlled non-temporal reference for evaluating frame-wise fusion quality, inter-frame output behavior, text conditioning, the clean-input effect of perturbation-based training, and downstream perception utility.

The main contributions are summarized as follows:

- We introduce UAV-VIVID, a large-scale UAV-acquired IR–VI video dataset containing 375 continuous scenes and 190,978 temporally paired frames, with scene-preserving organization and scene-disjoint partitions.
- We release registered lower-altitude fusion sequences, an unregistered challenge subset, higher-altitude auxiliary sequences, and structured frame-level descriptions. These task-oriented resources support video fusion, registration-uncertainty studies, and aerial perception.
- We develop the Residual Text-Guided Fusion Network (RT-FuseNet) as a text-guided frame-level baseline with Adaptive Residual Balancing and Text-Affine and Gated Modulation. Experiments evaluate frame-wise fusion quality, sequence-level output behavior, text conditioning, and downstream perception across multiple datasets.



Figure 2: Representative paired IR–VI scenes in UAV-VIVID across altitudes, illumination, weather, environments, crowds, and small or occluded targets.

Related Work

We review image-level IR–VI fusion, text-guided fusion, and video-level multimodal fusion.

Image-Level and Adversarial IR–VI Fusion

Image-level IR–VI fusion directly maps paired inputs to a fused image. Existing methods explore task-aware semantics Chen et al. (2025), high-order cross-modal interactions Zheng et al. (2024), degradation-aware modeling Wang et al. (2025a), Transformer and diffusion architectures Tang et al. (2024); Yue et al. (2023), and task-driven objectives Liu et al. (2025); Wu et al. (2025); Bai et al. (2025). Robust fusion has also been studied through PGD-based training and anti-attack losses Li et al. (2025a), perception-aware robustness Liu et al. (2023b), degradation modeling Wang et al. (2025a); Chen et al. (2025), and language-guided expert routing Li et al. (2026a). However, their output behavior on continuous UAV sequences remains less systematically characterized.

Text-Guided Fusion

Text-guided methods introduce language semantics into fusion. Text-IF Yi et al. (2024) supports interactive guidance, IF-FILM Zhao et al. (2024) combines generated descriptions with cross-attention, and LDFusion Wang et al. (2025b) defines language-based objectives in the CLIP Radford et al. (2021) space. UP-Fusion Li et al. (2026d) applies text-guided feature perturbation, while CtrlFuse Sun et al. (2026), RIS-Fuse Wang et al. (2025c), AWM-Fuse Li et al. (2026e), and ControlFuse Zhao et al. (2026) extend guidance to spatial, instance-level, adverse-weather, and instruction-based control. These methods are mainly evaluated on individual pairs, whereas UAV-VIVID enables analysis of language-conditioned outputs across UAV sequences.

Video-Level Fusion and Aerial Multimodal Benchmarks

Recent methods extend multimodal fusion to continuous video. AerialFusion Qiu et al. (2026) jointly addresses aerial registration and fusion, while VideoFusion Tang et al. (2026), UniVF Zhao et al. (2025), TemCoCo Gong et al. (2025), FTPFusion Li et al. (2026b), and MAVFusion Li et al. (2026c) investigate multi-frame interaction and temporal consistency. UniVF Zhao et al. (2025) further evaluates both spatial quality and temporal behavior.

Visible–thermal datasets such as LasHeR Li et al. (2022), DRGBT603 Ding et al. (2026), and VTUAV Zhang et al. (2022) primarily support tracking, while DroneVehicle Sun et al. (2022) targets

aerial vehicle detection. In contrast, UAV-VIVID is organized for scene-level UAV IR–VI fusion evaluation and provides low- and high-altitude sequences, registered fusion data, an unregistered challenge set, and structured textual descriptions.

THE UAV-VIVID BENCHMARK

COLLECTION, PAIRING, AND REGISTRATION

UAV-VIVID contains 190,978 temporally ordered IR–VI frame pairs from 375 UAV scene segments. It comprises three mutually exclusive subsets: higher-altitude auxiliary sequences, registration-processed lower-altitude fusion sequences, and an unregistered lower-altitude daytime challenge set. As summarized in Table 1, existing video-level IR–VI benchmarks differ substantially in their acquisition platforms, sequence organization, and alignment resources. Compared with the listed benchmarks, UAV-VIVID combines native low-altitude UAV acquisition, the largest number of scene segments, two explicitly separated altitude regimes, both registration-processed and raw cross-modal pairs, and structured frame-level text annotations. This organization enables controlled study of altitude-induced scale variation, registration uncertainty, and temporal behavior in continuous UAV IR–VI streams.

The data were captured using co-mounted infrared and visible sensors on a DJI Matrice 4T at a nominal 30 fps, with native resolutions of 1280×1024 and 3840×2160, respectively. Lower- and higher-altitude sequences were acquired below 20 m and at 20–40 m, respectively. The former provide larger targets and finer details for pixel-level fusion, whereas the latter offer wider views and smaller targets for aerial perception.

Scene segments are separated by changes in location, altitude, camera motion, illumination, or environment. After removing near-duplicate infrared frames, sequences with reliable DJI subtitle timestamps are paired by greedy one-to-one matching anchored on the infrared stream; the remaining sequences are paired by decoded frame order. Unmatched frames are discarded, while source indices and pairing metadata are retained.

Geometric calibration uses more than 5,000 paired IR–VI checkerboard images with diverse poses to estimate sensor intrinsics, distortion, and relative pose. For the lower-altitude fusion subset, verified cross-modal correspondences on representative keyframes are used to estimate homographies that warp visible frames onto the infrared image plane.

SCENE DIVERSITY AND TEXT ANNOTATIONS

As illustrated in Fig. 2, the dataset spans a broad range of UAV scenes and challenging conditions, including campus and urban environments, waterfront and greenery areas, infrastructure sites, dense crowds, rainy weather, low-light and reflective scenes, traffic intersections, and scenes containing small or partially occluded targets.

Each frame pair is accompanied by a fixed structured description generated offline. BLIP-2 Li et al. (2023) provides a global description of the visible scene, while YOLOv8 Jocher et al. (2023) identifies object categories and their spatial locations. Their outputs are combined to provide both scene-level context and object-level semantic cues.

BENCHMARK PROTOCOL

All frames from the same continuous scene are assigned to a single split to avoid temporal overlap. The registered lower-altitude benchmark contains 129 training scenes and 70 held-out test scenes. For model development, the training set is divided into 109 training and 20 scene-disjoint validation scenes. Model design, hyperparameter selection, checkpoint selection, and ablation studies are conducted only on this fixed partition. After the configuration is finalized, the model is retrained from scratch on all 129 training scenes and evaluated on the test set for final comparison. The higher-altitude and unregistered subsets are not used in these experiments.

Table 1: Comparison of representative continuous infrared–visible benchmarks relevant to video fusion. "Native" denotes data directly acquired using paired sensors, whereas "curated" denotes sequences selected from an existing dataset. Avg. Len. denotes the average number of paired frames per released sequence. A dash indicates that the corresponding quantity is not consistently specified by the canonical source.

Dataset	Primary Use	Source / Platform	Seq.	Pairs	Avg. Len.	Altitude	Alignment Resource	Structured Text
INO INO (2026)	Video analytics	Native / ground	15	–	–	–	Registered	–
HDO Xie et al. (2024); Xie (2024)	Registration + fusion	Native / ground	–	5,788	–	–	Raw + aligned	–
VF-Bench (IVF) Zhao et al. (2025)	Video fusion	Curated / VT-MOT	90	∼27K	∼300	–	Curated aligned	–
M^{3}SVD Tang et al. (2026)	Video fusion	Native / tripod	220	153,797	∼699	–	Registered	–
EUM3D Qiu et al. (2026)	Registration + fusion	Native / UAV	60	∼122K	∼2,033	100–500 m	Raw + transform GT	–
UAV-VIVID	**Video fusion + challenges**	**Native / UAV**	**375**	**190,978**	**∼509**	**∼20–40 m**	**Registered + raw subset**	**Yes**

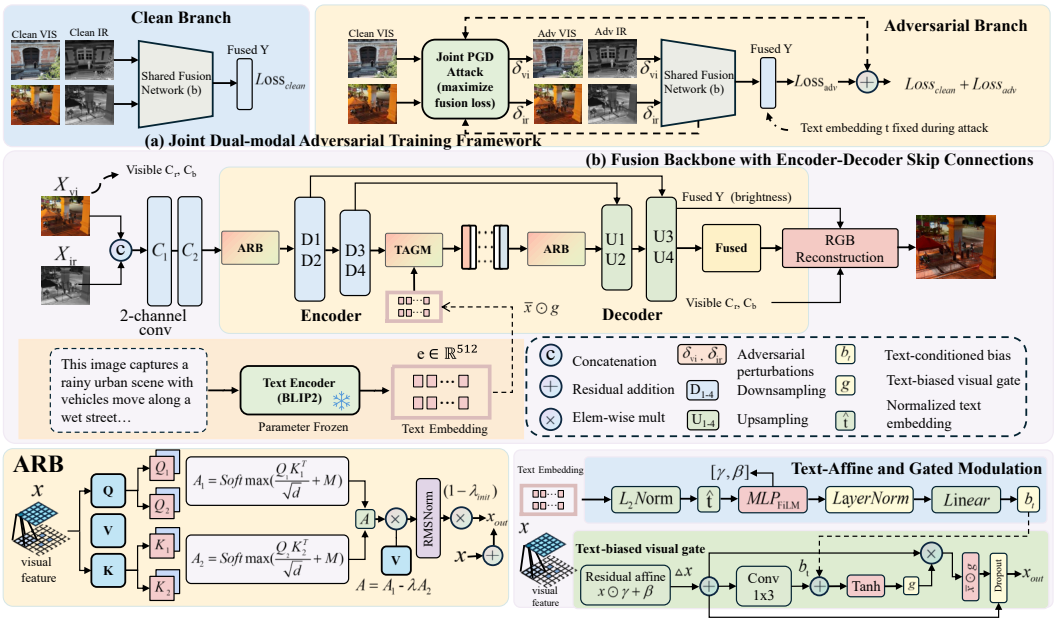


Figure 3: Architecture of RT-FuseNet. (a) Joint dual-modal PGD training with shared fusion parameters and a fixed text embedding during perturbation generation. (b) The encoder–decoder fusion backbone with ARB, TAGM, and multi-scale skip connections. Bottom: detailed structures of ARB and TAGM.

RT-FuseNet: A Frame-Level Reference Baseline

Overview

As illustrated in Fig. 3, RT-FuseNet comprises an encoder–decoder fusion backbone, two Adaptive Residual Balancing (ARB) modules, Text-Affine and Gated Modulation (TAGM), and a joint dual-modal PGD training framework.

Given a visible image X_{vi} and an infrared image X_{ir}, we convert X_{vi} into the YCrCb color space and concatenate its luminance channel with the infrared input:

$$X = \mathrm{Concat}\left(Y_{\mathrm{vi}}, X_{\mathrm{ir}}\right), \quad X \in \mathbb{R}^{B\times 2\times H\times W}. \tag{1}$$

Two convolutional layers first extract shallow features, followed by the first ARB module. The four-stage encoder then produces multi-scale representations, with TAGM injecting textual guidance into the deep features. A second ARB module refines the bottleneck representation before decoding, while encoder–decoder skip connections preserve spatial details at corresponding scales. The decoder predicts fused luminance, which is combined with the visible chrominance channels C_r and C_b to reconstruct the final RGB image.

Each description is encoded by a frozen BLIP-2 encoder Li et al. (2023) and mapped to a 512-dimensional embedding. During training, the clean and perturbed inputs share the same fusion network, while the text embedding remains fixed during perturbation generation. At inference time, RT-FuseNet consumes the IR–VI pair together with its offline-generated description.

ARB: Adaptive Residual Balancing

ARB regulates feature interactions through patch-level differential attention. Given $X \in \mathbb{R}^{B\times C\times H\times W}$, a non-overlapping 16×16 patch embedding produces $z \in \mathbb{R}^{B\times N\times C}$, where $N = \lfloor H/16 \rfloor \lfloor W/16 \rfloor$. For a 640×480 input, the two ARB modules operate on 1,200 and 70 tokens, respectively, rather than on the original pixel grid. We then compute

$$\begin{aligned} Q &= zW_Q = [Q_1 \parallel Q_2], \\ K &= zW_K = [K_1 \parallel K_2], \\ V &= zW_V, \end{aligned} \tag{2}$$

where $Q_i, K_i \in \mathbb{R}^{B\times N\times d}$, $d = C/2$, and $\parallel$ denotes channel-wise concatenation.

The two query–key subspaces produce separate attention maps, which are combined through differential subtraction:

$$\begin{aligned} A_i &= \mathrm{Softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d}} + M\right), \quad i \in \{1, 2\}, \\ A &= A_1 - \lambda_{\mathrm{dyn}} A_2. \end{aligned} \tag{3}$$

ARB uses global spatial attention; therefore, all token interactions are permitted and $M = \mathbf{0}_{N\times N}$.

Following differential attention Ye et al. (2025), the subtractive coefficient is parameterized as

$$\begin{aligned} \lambda_{\mathrm{dyn}} = \; & \exp(\langle \boldsymbol{\lambda}_{q1}, \boldsymbol{\lambda}_{k1} \rangle) \\ & - \exp(\langle \boldsymbol{\lambda}_{q2}, \boldsymbol{\lambda}_{k2} \rangle) + \lambda_{\mathrm{init}}, \end{aligned} \tag{4}$$

where the four $\boldsymbol{\lambda}$ vectors are learnable. The depth-dependent initialization term is fixed as

$$\lambda_{\mathrm{init}} = 0.8 - 0.6 \exp[-0.3(\ell - 1)], \tag{5}$$

where ℓ denotes the block depth. The balanced response is aggregated through V and residually injected:

$$z_{\mathrm{out}} = z + (1 - \lambda_{\mathrm{init}}) \, \mathrm{RMSNorm}(AV). \tag{6}$$

After restoring the spatial layout, the subtractive branch suppresses redundant interactions while preserving complementary cross-modal information.

TAGM: Text-Affine and Gated Modulation

TAGM converts textual semantics into FiLM-style channel-wise affine parameters Perez et al. (2018) and a text-conditioned visual gate. For a description p, a frozen BLIP-2 encoder and a lightweight projection produce

$$\begin{aligned} h &= E_{\mathrm{BLIP2}}(p), \\ e &= W_{\mathrm{proj}} h, \qquad \hat{t} = \frac{e}{\|e\|_2}, \end{aligned} \tag{7}$$

where $e \in \mathbb{R}^{B\times 512}$ and $\hat{t}$ is its normalized representation.

The normalized embedding is transformed into channel-wise residual affine parameters and a text-conditioned bias:

$$\begin{aligned} u &= \mathrm{MLP}_{\mathrm{FiLM}}(\hat{t}) = [\gamma \parallel \beta], \\ b_t &= \mathrm{Linear}_b \left(\mathrm{LayerNorm}(u) \right), \end{aligned} \tag{8}$$

where $\gamma, \beta, b_t \in \mathbb{R}^{B\times C}$.

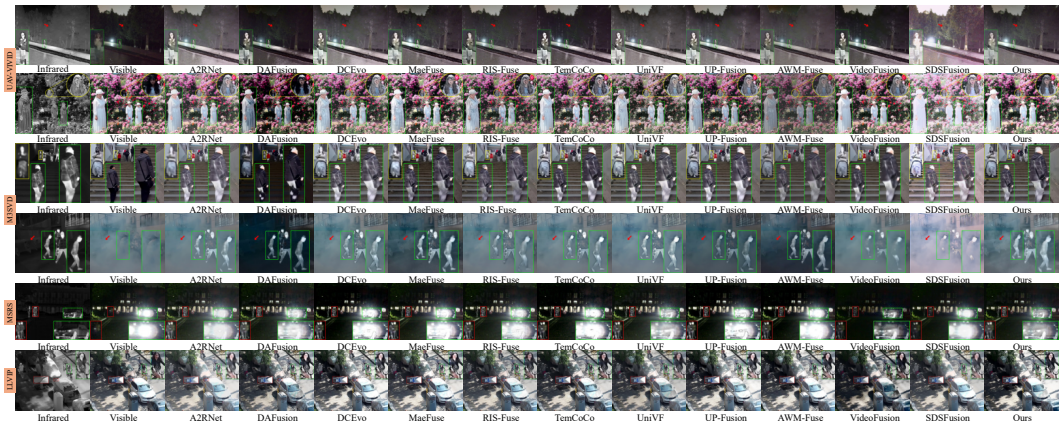


Figure 4: Qualitative comparison on LLVIP, MSRS, M^{3}SVD Tang et al. (2026), and UAV-VIVID, showing thermal-target, visible-detail, and structure preservation.

For a visual feature $x \in \mathbb{R}^{B\times C\times H\times W}$, TAGM first predicts a text-conditioned residual affine update and then applies a spatially adaptive gate:

$$\begin{aligned} \Delta x &= x \odot \gamma^{\uparrow} + \beta^{\uparrow}, \\ \bar{x} &= x + \Delta x, \\ g &= \tanh\Big(\mathcal{C}_{1\times 3}(\bar{x}) + b_t^{\uparrow}\Big), \\ x_{\text{out}} &= \text{Dropout}(\bar{x} + \bar{x} \odot g). \end{aligned} \tag{9}$$

Here, $\uparrow$ denotes spatial broadcasting, $\odot$ denotes element-wise multiplication, and $\mathcal{C}_{1\times 3}$ is a 2D convolution operating along the spatial width dimension. The residual affine branch produces the text-conditioned update Δx, while the visual gate provides spatially adaptive modulation. When $\gamma = \beta = 0$, we have $\Delta x = 0$ and $\bar{x} = x$. If the gate response is also close to zero, TAGM reduces to an identity mapping at inference, where dropout is disabled. Therefore, weak textual modulation preserves the original visual representation.

Joint Dual-Modal PGD Training

As shown in Fig. 3(a), joint PGD Madry et al. (2018) perturbs the visible and infrared inputs simultaneously during training. Supervision is derived exclusively from the clean inputs. We first construct the intensity and gradient targets:

$$\begin{aligned} T_{\text{int}} &= \max(Y_{\text{vi}}, X_{\text{ir}}), \\ T_{\text{grad}} &= \max\left(\mathcal{G}(Y_{\text{vi}}), \mathcal{G}(X_{\text{ir}})\right), \end{aligned} \tag{10}$$

where $\mathcal{G}$ denotes gradient magnitude and both maxima are computed element-wise.

For fused luminance I, the source-derived losses are

$$\begin{aligned} \mathcal{L}_{\text{int}}(I) &= \frac{\|I - T_{\text{int}}\|_1}{BHW}, \\ \mathcal{L}_{\text{grad}}(I) &= \frac{\|\mathcal{G}(I) - T_{\text{grad}}\|_1}{BHW}, \\ \mathcal{L}_{\text{str}}(I) &= 1 - \frac{1}{2}\Big[\text{SSIM}(I, Y_{\text{vi}}) + \text{SSIM}(I, X_{\text{ir}})\Big]. \end{aligned} \tag{11}$$

The fusion objective combines the three terms:

$$\begin{aligned} \mathcal{L}_{\text{fus}}(I) = \lambda_{\text{int}}\mathcal{L}_{\text{int}}(I) + \lambda_{\text{grad}}\mathcal{L}_{\text{grad}}(I) \\ + \lambda_{\text{str}}\mathcal{L}_{\text{str}}(I). \end{aligned} \tag{12}$$

Methods [Pub./Year]	Dataset: UAV-VIVID							Dataset: M^{3}SVD							Dataset: MSRS					Dataset: LLVIP				
	BiSWE↓	flowD↓	SEN↑	STD↑	SSIM↑	SCD↑	NQM↑	BiSWE↓	flowD↓	SEN↑	STD↑	SSIM↑	SCD↑	NQM↑	SEN↑	STD↑	SSIM↑	SCD↑	NQM↑	SEN↑	STD↑	SSIM↑	SCD↑	NQM↑
A2RNet [AAAI'25]	3.921	4.529	7.404	50.329	0.628	1.266	4.372	2.771	3.398	6.943	43.511	0.581	1.543	1.831	6.929	46.814	0.765	0.543	13.703	7.010	40.650	0.588	1.374	3.699
DAFusion [IF'25]	5.487	5.825	7.253	56.699	0.592	1.419	3.889	3.150	3.585	6.180	39.247	0.522	1.410	1.024	6.576	46.910	0.743	0.716	12.076	5.923	26.609	0.561	1.082	3.722
DCEvo [CVPR'25]	7.262	7.602	7.598	56.309	0.642	1.373	3.912	4.103	4.542	7.341	53.759	0.641	1.751	1.241	6.977	46.376	0.825	0.342	12.137	7.096	43.230	0.661	1.524	3.390
MaeFuse [TIP'25]	5.738	6.057	7.638	58.399	0.620	1.621	3.979	4.575	4.988	7.180	42.422	0.618	1.770	1.612	7.101	**54.774**	0.793	1.393	9.712	6.947	40.278	0.644	1.630	2.896
RIS-Fuse [ICCV'25]	7.874	8.208	7.528	53.337	0.627	1.105	4.558	4.665	5.183	7.318	51.592	0.632	1.618	2.272	6.932	45.820	0.842	0.538	12.961	6.882	38.613	0.659	1.375	2.717
TemCoCo [ICCV'25]	8.962	10.038	7.528	56.901	0.619	1.232	4.456	4.238	5.068	7.295	54.799	0.620	1.618	1.636	6.658	48.537	0.744	0.588	9.611	6.566	35.995	0.600	1.178	2.363
UniVF [NeurIPS'25]	6.842	7.304	7.611	56.530	0.638	1.399	4.111	4.071	4.749	7.258	53.003	0.649	1.809	1.312	7.026	51.106	0.823	1.189	12.389	7.117	44.458	0.647	1.679	3.274
UP-Fusion [AAAI'26]	9.670	12.302	7.455	51.309	0.654	1.175	4.016	7.856	8.523	7.227	50.058	0.658	1.722	1.751	6.921	46.172	0.851	0.466	12.402	6.963	41.419	0.676	1.439	3.499
AWM-Fuse [TIP'26]	5.851	5.926	6.891	33.186	0.630	0.577	4.049	5.449	5.630	6.803	35.246	0.652	1.478	2.297	6.637	36.514	0.843	0.725	10.862	6.483	29.169	0.640	1.025	**3.903**
VideoFusion [CVPR'26]	**3.777**	**4.329**	7.494	55.899	0.633	1.223	4.193	**2.486**	**3.167**	7.214	54.320	0.647	1.702	1.140	6.455	30.192	0.631	1.070	5.202	6.228	24.185	0.646	0.636	2.163
SDSFusion [TIP'25]	7.160	7.433	7.583	**61.800**	0.623	1.633	4.362	6.569	7.175	7.581	**60.084**	0.454	1.329	1.205	7.194	52.306	0.562	0.567	5.699	**7.293**	**47.502**	0.328	1.149	0.200
Ours	5.287	7.277	**7.719**	60.597	**0.678**	**1.655**	**4.579**	3.385	4.055	**7.682**	59.586	**0.665**	**1.855**	**2.337**	**7.255**	53.895	**0.885**	**1.455**	**13.951**	7.249	45.693	**0.687**	**1.753**	3.724

Table 2: Within each evaluation track, all methods are retrained from their official implementations using identical source training data and evaluated on the same 640×480 test lists without target-domain fine-tuning. The UAV-VIVID-trained models are evaluated on UAV-VIVID and M3SVD, whereas the FMB-trained models are evaluated on MSRS and LLVIP.

Based on the initial magnitudes of the training losses, we set $(\lambda_{\text{int}}, \lambda_{\text{grad}}, \lambda_{\text{str}}) = (1, 10, 1)$.

The perturbations are generated by maximizing the same fusion objective:

$$(\delta_{\text{vi}}^*, \delta_{\text{ir}}^*) = \arg \max_{\max\{\|\delta_{\text{vi}}\|_\infty, \|\delta_{\text{ir}}\|_\infty\} \le \epsilon} \mathcal{L}_{\text{fus}}(I_f^{\text{adv}}). \tag{13}$$

The clean-source targets and normalized text embedding $\hat{t}$ are held fixed during perturbation generation. The clean and perturbed outputs are then optimized against the same supervision:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fus}}(I_f) + \mathcal{L}_{\text{fus}}(I_f^{\text{adv}}). \tag{14}$$

EXPERIMENTS

We first describe the experimental setup and evaluate the proposed method in terms of fusion performance, downstream object detection, and key design choices. The supplementary material provides dataset release splits and the development protocol, full statistics for the main comparisons and ablations, computational-efficiency and semantic- segmentation results, additional FMB and M3FD comparisons, and details on data governance, temporal synchronization, geometric calibration, and residual misalignment.

EXPERIMENTAL SETUP

Datasets and evaluation protocols. For UAV-VIVID, model development and all ablations use a fixed 109/20 scene-level training/validation split. The supplementary ablations report validation results. After all settings are fixed, RT-FuseNet is retrained on all 129 training scenes and evaluated on the 70-scene test set, as reported in Table 2.

The resulting model is also directly evaluated on M^{3}SVD Tang et al. (2026) without using its training set. For image-level evaluation, a separate model is trained on FMB Liu et al. (2023a) and tested on FMB, MSRS Tang et al. (2022), LLVIP Jia et al. (2021), and M3FD Liu et al. (2022) without fine-tuning. For downstream tasks, only the target-domain detector or segmenter is trained, with results reported on LLVIP and MSRS, respectively.

Implementation details. RT-FuseNet is trained for 50 epochs on one 48-GB NVIDIA L20 GPU using AdamW with a batch size of 2, an initial learning rate of 5×10^{-5}, a weight decay of 10^{-4}, and cosine annealing to 10^{-8}. During development, the checkpoint with the lowest source-derived fusion objective in Eq. (12) on the fixed validation split is selected. For final evaluation, the finalized configuration is retrained for 50 epochs on the full development set, and the last checkpoint is used.

Three-step PGD Madry et al. (2018) jointly perturbs both modalities with $\epsilon = 4/255$ under the ℓ_∞ norm and a step size of $1/255$. The intensity, gradient, and structural supervision terms are derived from clean inputs and held fixed during perturbation generation. For downstream evaluation, YOLOv8s Jocher et al. (2023) is trained for 100 epochs with 640×640 inputs and a batch size of 64, while semantic segmentation follows SegMiF Liu et al. (2023a).

Evaluation Metrics and Compared Methods

We evaluate sequence-level output consistency using BiSWE Zhao et al. (2025) and flowD Gong et al. (2025), where lower values indicate smaller inter-frame variation. Frame-wise fusion quality is assessed using Shannon entropy (SEN) Shannon (1948), standard deviation (STD) Sun et al. (2023), structural similarity (SSIM) Wang et al. (2004), the sum of correlations of differences (SCD) Aslantas & Bendes (2015), and the noise quality measure (NQM) Damera-Venkata et al. (2000). These complementary metrics characterize information richness, contrast, structural preservation, source-information transfer, and perceptual quality, respectively. Frame-wise scores are averaged over all evaluated frames, whereas BiSWE and flowD are computed over valid consecutive frame pairs within each evaluated scene.

Ablation summary. The full prompt and ARB/TAGM ablations are provided in the supplementary material; the joint configuration achieves the best validation trade-off across the reported metrics.

Fusion Performance

Qualitative comparison. As shown in Fig. 4, RT-FuseNet consistently preserves salient thermal targets while retaining visible textures and clear object boundaries. ARB regulates redundant or unreliable cross-modal interactions, whereas TAGM uses textual semantics to adaptively modulate channel responses and spatial regions. Together with the source-derived intensity, gradient, and structural objectives, these components produce fused images with well-preserved target saliency, local detail, and structural continuity.

Quantitative comparison. As shown in Table 2, sequence-aware methods such as VideoFusion benefit from neighboring frames and achieve stronger temporal metrics. Despite processing frames independently, RT-FuseNet remains competitive, ranking third in BiSWE on UAV-VIVID and fourth in both BiSWE and flowD on M^3SVD. It shows clearer advantages in frame-wise quality, achieving the best SEN, SSIM, SCD, and NQM on UAV-VIVID, M^3SVD, and MSRS, as well as the best SSIM and SCD on LLVIP. These results support the effectiveness of ARB and TAGM in cross-modal interaction and text-conditioned modulation.

Detailed downstream detection and segmentation results are provided in the supplementary material. On LLVIP, RT-FuseNet reaches 0.8598 precision and 0.6425 $AP_{50:95}$ under the shared YOLOv8s protocol.

Conclusion

We introduce UAV-VIVID, a large-scale UAV IR–VI video dataset with 375 temporally ordered scenes and 190,978 paired frames, together with RT-FuseNet, a text-guided frame-level reference baseline. ARB regulates cross-modal interactions, while TAGM incorporates description-derived guidance. Experiments demonstrate strong frame-wise fusion quality, competitive sequence-level output consistency, and downstream utility. Future work will extend the benchmark toward explicit temporal modeling, motion-aware evaluation, and more general cross-modal registration.

AI use statement

We used generative AI tools to assist with LaTeX formatting and manuscript file migration. They were not used to generate the research ideas, data, experimental results, or scientific claims. The authors are responsible for reviewing AI-assisted formatting changes and for the final content of this work.

References

Veysel Aslantas and Emre Bendes. A new image quality metric for image fusion: The sum of the correlations of differences. *AEU - International Journal of Electronics and Communications*, 69 (12):1890–1896, 2015. doi: 10.1016/j.aeue.2015.09.004.

Haowen Bai, Jiangshe Zhang, Zixiang Zhao, Yichen Wu, Lilun Deng, Yukun Cui, Tao Feng, and Shuang Xu. Task-driven image fusion with learnable fusion loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7457–7468, June 2025.

Jun Chen, Liling Yang, Wei Yu, Wenping Gong, Zhanchuan Cai, and Jiayi Ma. Sdsfusion: A semantic-aware infrared and visible image fusion network for degraded scenes. *IEEE Transactions on Image Processing*, 34:3139–3153, 2025. doi: 10.1109/TIP.2025.3571339.

Niranjan Damera-Venkata, Thomas D. Kite, Wilson S. Geisler, Brian L. Evans, and Alan C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4):636–650, 2000. doi: 10.1109/83.841940.

Zhaodong Ding, Chenglong Li, Shengqing Miao, and Jin Tang. Causality-based modality- and platform-invariant representation learning for dynamic rgbt tracking and a benchmark. *IEEE Transactions on Image Processing*, 35:3141–3156, 2026. doi: 10.1109/TIP.2026.3674367.

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 (ICCV)*, pp. 14326–14335, October 2025. doi: 10.1109/ICCV51701.2025.01329.

INO. INO Video Analytics Dataset. `https://www.ino.ca/en/technologies/video-analytics-dataset/videos/`, 2026. Accessed: 2026-07-29.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 3496–3504, October 2021. doi: 10.1109/ICCVW54120.2021.00389.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8. `https://github.com/ultralytics/ultralytics`, 2023. Version 8.0.0. Accessed: 2026-07-29.

Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. Lasher: A large-scale high-diversity benchmark for rgbt tracking. *IEEE Transactions on Image Processing*, 31:392–404, 2022. doi: 10.1109/TIP.2021.3130533.

Jiawei Li, Hongwei Yu, Jiansheng Chen, Xinlong Ding, Jinlong Wang, Jinyuan Liu, Bochao Zou, and Huimin Ma. A^2rnet: Adversarial attack resilient network for robust infrared and visible image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4770–4778, 2025a. doi: 10.1609/aaai.v39i5.32504.

Jiayang Li, Junjun Jiang, Pengwei Liang, Jiayi Ma, and Liqiang Nie. Maefuse: Transferring omni features with pretrained masked autoencoders for infrared and visible image fusion via guided training. *IEEE Transactions on Image Processing*, 34:1340–1353, 2025b. doi: 10.1109/TIP.2025.3541562.

Jing Li, Yifan Wang, Jiafeng Yan, Renlong Zhang, and Bin Yang. MdaIF: Robust one-stop multi-degradation-aware image fusion with language-driven semantics. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6226–6234, 2026a. doi: 10.1609/aaai.v40i8.37548.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/li23q.html`.

Xilai Li, Chusheng Fang, and Xiaosong Li. Ftpfusion: Frequency-aware infrared and visible video fusion with temporal perturbation, 2026b.

Xilai Li, Weijun Jiang, Xiaosong Li, Yang Liu, Hongbin Wang, Tao Ye, Huafeng Li, and Haishu Tan. Mavfusion: Efficient infrared and visible video fusion via motion-aware sparse interaction. In *European Conference on Computer Vision (ECCV)*, 2026c.

Xilai Li, Xiaosong Li, and Weijun Jiang. Text-guided channel perturbation and pre-trained knowledge integration for unified multi-modality image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6521–6529, 2026d. doi: 10.1609/aaai.v40i8.37581.

Xilai Li, Huichun Liu, Xiaosong Li, Tao Ye, Zhenyu Kuang, and Huafeng Li. AWM-Fuse: Multi-modality image fusion for adverse weather via global and local text perception. *IEEE Transactions on Image Processing*, 35:5151–5164, 2026e. doi: 10.1109/TIP.2026.3690324.

Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5802–5811, June 2022.

Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8115–8124, October 2023a. doi: 10.1109/ICCV51070.2023.00745.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2226–2235, June 2025. doi: 10.1109/CVPR52734.2025.00213.

Zhu Liu, Jinyuan Liu, Benzhuang Zhang, Long Ma, Xin Fan, and Risheng Liu. PAIF: Perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3706–3714. ACM, 2023b. doi: 10.1145/3581783.3611928.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

Junhui Qiu, Xiang Xiang, Hongyun Wang, and Jiaqi Gui. Aerialfusion: Co-motion-driven unified registration and fusion on multi-modal data streams from aerial view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8583–8591, 2026. doi: 10.1609/aaai.v40i10.37810.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 623–656, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Bin Sun, Yunxiang Gao, Wuwei Zhuge, and Zixuan Wang. Analysis of quality objective assessment metrics for visible and infrared image fusion. *Journal of Image and Graphics*, 28(1):144–155, 2023. doi: 10.11834/jig.210719.

Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6700–6713, 2022. doi: 10.1109/TCSVT.2022.3168279.

Yiming Sun, Yuan Ruan, Qinghua Hu, and Pengfei Zhu. Ctrlfuse: Mask-prompt guided controllable infrared and visible image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9260–9268, 2026. doi: 10.1609/aaai.v40i11.37884.

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*, 83–84:79–92, 2022. doi: 10.1016/j.inffus.2022.03.007.

Linfeng Tang, Yeda Wang, Meiqi Gong, Zizhuo Li, Yuxin Deng, Xunpeng Yi, Chunyu Li, Han Xu, Hao Zhang, and Jiayi Ma. Videofusion: A spatio-temporal collaborative network for multi-modal video fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19559–19569, June 2026.

Wei Tang, Fazhi He, and Yu Liu. Itfuse: An interactive transformer for infrared and visible image fusion. *Pattern Recognition*, 156:110822, 2024. doi: 10.1016/j.patcog.2024.110822.

Xue Wang, Zheng Guan, Wenhua Qian, Jinde Cao, Runzhuo Ma, and Cong Bi. A degradation-aware guided fusion network for infrared and visible image. *Information Fusion*, 118:102931, 2025a. doi: 10.1016/j.inffus.2025.102931.

Yuhao Wang, Lingjuan Miao, Zhiqiang Zhou, Lei Zhang, and Yajun Qiao. Infrared and visible image fusion with language-driven loss in CLIP embedding space. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 1443–1451. ACM, 2025b. doi: 10.1145/3746027.3755182.

Zeyu Wang, Jizheng Zhang, Haiyu Song, Mingyu Ge, Jiayu Wang, and Haoran Duan. Highlight what you want: Weakly-supervised instance-level controllable infrared-visible image fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12637–12647, October 2025c. doi: 10.1109/ICCV51701.2025.01174.

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. doi: 10.1109/TIP.2003.819861.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17882–17891, June 2025.

Housheng Xie. HDO: A high-quality dual-optical video sequence dataset. `https://github.com/xiehousheng/HDO`, 2024. Accessed: 2026-07-29.

Housheng Xie, Meng Sang, Yukuan Zhang, Yang Yang, Shan Zhao, and Jianbo Zhong. RCVS: A unified registration and fusion framework for video streams. *IEEE Transactions on Multimedia*, 26:11031–11043, 2024. doi: 10.1109/TMM.2024.3443673.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=OvoCm1gGhN`.

Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Text-if: Leveraging semantic text guidance for degradation-aware and interactive image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27026–27035, June 2024. doi: 10.1109/CVPR52733.2024.02552.

Jun Yue, Leyuan Fang, Shaobo Xia, Yue Deng, and Jiayi 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. doi: 10.1109/TIP.2023.3322046.

Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8876–8885, June 2022. doi: 10.1109/CVPR52688.2022.00868.

Libo Zhao, Xiaoli Zhang, and Zeyu Wang. ControlFuse: Instruction-guided multi-granularity controllable image fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 13199–13207, 2026. doi: 10.1609/aaai.v40i16.38321.

Zixiang Zhao, Lilun Deng, Haowen Bai, Yukun Cui, Zhipeng Zhang, Yulun Zhang, Haotong Qin, Dongdong Chen, Jiangshe Zhang, Peng Wang, and Luc Van Gool. Image fusion via vision-language model. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60749–60765. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/zhao24d.html.

Zixiang Zhao, Haowen Bai, Bingxin Ke, Yukun Cui, Lilun Deng, Yulun Zhang, Kai Zhang, and Konrad Schindler. A unified solution to video fusion: From multi-frame learning to benchmarking. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 128504–128535. Curran Associates, Inc., 2025.

Naishan Zheng, Man Zhou, Jie Huang, Junming Hou, Haoying Li, Yuan Xu, and Feng Zhao. Probing synergistic high-order interaction in infrared and visible image fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26374–26385, June 2024. doi: 10.1109/CVPR52733.2024.02492.

1 Additional Experimental Results

Scene-Segment Release Splits

As summarized in Table 3, UAV-VIVID contains 375 scene segments and 190,978 IR–VI frame pairs. The higher-altitude subset comprises 78 training, 19 validation, and 48 test scenes, totaling 79,992 pairs. The registration-processed lower-altitude subset contains 129 training scenes with 61,475 pairs and 70 test scenes with 34,612 pairs. In addition, 31 unregistered daytime scenes containing 14,899 pairs are released as a separate challenge set.

All UAV-VIVID model development, ablation, and final evaluation use only the registration-processed lower-altitude subset. For model development, 20 scenes are held out from its 129-scene training split as a scene-disjoint validation set, leaving 109 scenes for development training. After the configuration is fixed, the final model is retrained from scratch on all 129 training scenes and evaluated on the 70-scene test split. The higher-altitude and unregistered subsets are not used in the reported UAV-VIVID fusion experiments.

Subset	**Split**	**Scenes**	**Pairs**
High-alt. sequences	Train	78	42,936
	Val	19	10,126
	Test	48	26,930
Low-alt. reg.-proc.	Train	129	61,475
	Test	70	34,612
Unreg. low-day	Challenge	31	14,899

Table 3: Scene-segment release splits. All UAV-VIVID model development, ablation, and final evaluation use only the registration-processed lower-altitude subset.

FULL CROSS-DATASET COMPARISON

The following table reports the complete cross-dataset comparison referenced by the main paper.

Methods [Pub./Year]	Dataset: UAV-VIVID							Dataset: M^{3}SVD							Dataset: MSRS					Dataset: LLVIP				
	BiSWE↓	flowD↓	SEN↑	STD↑	SSIM↑	SCD↑	NQM↑	BiSWE↓	flowD↓	SEN↑	STD↑	SSIM↑	SCD↑	NQM↑	SEN↑	STD↑	SSIM↑	SCD↑	NQM↑	SEN↑	STD↑	SSIM↑	SCD↑	NQM↑
A2RNet [AAAI'25]	3.921	4.529	7.404	50.329	0.628	1.266	4.372	2.771	3.398	6.943	43.511	0.581	1.543	1.831	6.929	46.814	0.765	0.543	13.703	7.010	40.650	0.588	1.374	3.699
DAFusion [IF'25]	5.487	5.825	7.253	56.699	0.592	1.419	3.889	3.150	3.585	6.180	39.247	0.522	1.410	1.024	6.576	46.910	0.743	0.716	12.076	5.923	26.609	0.561	1.082	3.722
DCEvo [CVPR'25]	7.262	7.602	7.598	56.309	0.642	1.373	3.912	4.103	4.542	7.341	53.759	0.641	1.751	1.241	6.977	46.376	0.825	0.342	12.137	7.096	43.230	0.661	1.524	3.390
MaeFuse [TIP'25]	5.738	6.057	7.638	58.399	0.620	1.621	3.979	4.575	4.988	7.180	42.422	0.618	1.770	1.612	7.101	**54.774**	0.793	1.393	9.712	6.947	40.278	0.644	1.630	2.896
RIS-Fuse [ICCV'25]	7.874	8.208	7.528	53.337	0.627	1.105	4.558	4.665	5.183	7.318	51.592	0.632	1.618	2.272	6.932	45.820	0.842	0.538	12.961	6.882	38.613	0.659	1.375	2.717
TemCoCo [ICCV'25]	8.962	10.038	7.528	56.901	0.619	1.232	4.456	4.238	5.068	7.295	54.799	0.620	1.618	1.636	6.658	48.537	0.744	0.588	9.611	6.566	35.995	0.600	1.178	2.363
UniVF [NeurIPS'25]	6.842	7.304	7.611	56.530	0.638	1.399	4.111	4.071	4.749	7.258	53.003	0.649	1.809	1.312	7.026	51.106	0.823	1.189	12.389	7.117	44.458	0.647	1.679	3.274
UP-Fusion [AAAI'26]	9.670	12.302	7.455	51.309	0.654	1.175	4.016	7.856	8.523	7.227	50.058	0.658	1.722	1.751	6.921	46.172	0.851	0.466	12.402	6.963	41.419	0.676	1.439	3.499
AWM-Fuse [TIP'26]	5.851	5.926	6.891	33.186	0.630	0.577	4.049	5.449	5.630	6.803	35.246	0.652	1.478	2.297	6.637	36.514	0.843	0.725	10.862	6.483	29.169	0.640	1.025	**3.903**
VideoFusion [CVPR'26]	**3.777**	**4.329**	7.494	55.899	0.633	1.223	4.193	**2.486**	**3.167**	7.214	54.320	0.647	1.702	1.140	6.455	30.192	0.631	1.070	5.202	6.228	24.185	0.646	0.636	2.163
SDSFusion [TIP'25]	7.160	7.433	7.583	**61.800**	0.623	1.633	4.362	6.569	7.175	7.581	**60.084**	0.454	1.329	1.205	7.194	52.306	0.562	0.567	5.699	**7.293**	**47.502**	0.328	1.149	0.200
Ours	5.287	7.277	**7.719**	60.597	**0.678**	**1.655**	**4.579**	3.385	4.055	**7.682**	59.586	**0.665**	**1.855**	**2.337**	**7.255**	53.895	**0.885**	**1.455**	**13.951**	7.249	45.693	**0.687**	**1.753**	3.724

Table 4: Within each evaluation track, all methods are retrained from their official implementations using identical source training data and evaluated on the same 640×480 test lists without target-domain fine-tuning. The UAV-VIVID-trained models are evaluated on UAV-VIVID and M3SVD, whereas the FMB-trained models are evaluated on MSRS and LLVIP.

FULL STATISTICS FOR THE MAIN COMPARISON

Tables 5, 6, 7, and 8 provide the complete mean-and-standard-deviation results corresponding to the mean-only main-comparison table in the paper. Frame-wise metrics are summarized over individual test frames, whereas BiSWE and flowD are computed within each ordered test scene and then aggregated across scenes.

Methods [Pub./Year]	BiSWE↓	flowD↓	SEN↑	STD↑	SSIM↑	SCD↑	NQM↑
A2RNet [AAAI'25]	3.921 ± 2.265	4.529 ± 3.121	7.404 ± 0.273	50.329 ± 8.409	0.628 ± 0.047	1.266 ± 0.246	4.372 ± 2.089
DAFusion [IF'25]	5.487 ± 3.029	5.825 ± 3.437	7.253 ± 0.282	56.699 ± 9.918	0.592 ± 0.044	1.419 ± 0.172	3.889 ± 1.748
DCEvo [CVPR'25]	7.262 ± 4.386	7.602 ± 4.910	7.598 ± 0.267	56.309 ± 9.022	0.642 ± 0.044	1.373 ± 0.213	3.912 ± 2.097
MaeFuse [TIP'25]	5.738 ± 3.109	6.057 ± 3.724	7.638 ± 0.230	58.399 ± 10.346	0.620 ± 0.047	1.621 ± 0.134	3.979 ± 1.622
RIS-Fuse [ICCV'25]	7.874 ± 4.581	8.208 ± 5.110	7.528 ± 0.284	53.337 ± 8.454	0.627 ± 0.049	1.105 ± 0.300	4.558 ± 1.930
TemCoCo [ICCV'25]	8.962 ± 4.399	10.038 ± 5.962	7.528 ± 0.328	56.901 ± 9.336	0.619 ± 0.044	1.232 ± 0.242	4.456 ± 2.326
UniVF [NeurIPS'25]	6.842 ± 4.269	7.304 ± 4.812	7.611 ± 0.248	56.530 ± 8.566	0.638 ± 0.046	1.399 ± 0.205	4.111 ± 2.132
UP-Fusion [AAAI'26]	9.670 ± 20.078	12.302 ± 20.836	7.455 ± 0.346	51.309 ± 9.858	0.654 ± 0.048	1.175 ± 0.362	4.016 ± 2.060
AWM-Fuse [TIP'26]	5.851 ± 3.072	5.926 ± 3.456	6.891 ± 0.181	33.186 ± 4.624	0.630 ± 0.055	0.577 ± 0.511	4.049 ± 1.414
VideoFusion [CVPR'26]	$\mathbf{3.777 \pm 2.665}$	$\mathbf{4.329 \pm 3.838}$	7.494 ± 0.328	55.899 ± 9.657	0.633 ± 0.047	1.223 ± 0.251	4.193 ± 2.206
SDSFusion [TIP'25]	7.160 ± 4.564	7.433 ± 5.319	7.583 ± 0.274	$\mathbf{61.800 \pm 11.649}$	0.623 ± 0.050	1.633 ± 0.111	4.362 ± 1.749
Ours	5.287 ± 2.758	7.277 ± 3.527	$\mathbf{7.719 \pm 0.258}$	60.597 ± 9.942	$\mathbf{0.678 \pm 0.049}$	$\mathbf{1.655 \pm 0.258}$	$\mathbf{4.579 \pm 1.643}$

Table 5: Quantitative comparison on UAV-VIVID under the unified sequence-level protocol. All methods are retrained from official implementations on the UAV-VIVID source split and tested without target-domain fine-tuning on the identical 640×480 test list using the same metrics. Results are reported as mean $\pm$ standard deviation.

Methods [Pub./Year]	BiSWE↓	flowD↓	SEN↑	STD↑	SSIM↑	SCD↑	NQM↑
A2RNet [AAAI'25]	2.771 ± 1.910	3.398 ± 3.300	6.943 ± 0.366	43.511 ± 9.760	0.581 ± 0.070	1.543 ± 0.188	1.831 ± 1.037
DAFusion [IF'25]	3.150 ± 1.948	3.585 ± 2.874	6.180 ± 0.371	39.247 ± 14.029	0.522 ± 0.072	1.410 ± 0.199	1.024 ± 1.933
DCEvo [CVPR'25]	4.103 ± 2.619	4.542 ± 4.233	7.341 ± 0.184	53.759 ± 10.447	0.641 ± 0.052	1.751 ± 0.134	1.241 ± 1.126
MaeFuse [TIP'25]	4.575 ± 2.491	4.988 ± 3.777	7.180 ± 0.243	42.422 ± 7.537	0.618 ± 0.063	1.770 ± 0.107	1.612 ± 0.673
RIS-Fuse [ICCV'25]	4.665 ± 2.696	5.183 ± 4.578	7.318 ± 0.226	51.592 ± 11.448	0.632 ± 0.056	1.618 ± 0.243	2.272 ± 1.444
TemCoCo [ICCV'25]	4.238 ± 2.670	5.068 ± 3.845	7.295 ± 0.176	54.799 ± 13.500	0.620 ± 0.054	1.618 ± 0.216	1.636 ± 1.378
UniVF [NeurIPS'25]	4.071 ± 2.552	4.749 ± 4.061	7.258 ± 0.223	53.003 ± 14.733	0.649 ± 0.055	1.809 ± 0.114	1.312 ± 1.075
UP-Fusion [AAAI'26]	7.856 ± 15.412	8.523 ± 15.894	7.227 ± 0.246	50.058 ± 13.849	0.658 ± 0.052	1.722 ± 0.173	1.751 ± 1.610
AWM-Fuse [TIP'26]	5.449 ± 2.819	5.630 ± 3.954	6.803 ± 0.224	35.246 ± 7.069	0.652 ± 0.060	1.478 ± 0.241	2.297 ± 1.008
VideoFusion [CVPR'26]	$\mathbf{2.486 \pm 1.701}$	$\mathbf{3.167 \pm 2.837}$	7.214 ± 0.236	54.320 ± 15.154	0.647 ± 0.056	1.702 ± 0.179	1.140 ± 1.135
SDSFusion [TIP'25]	6.569 ± 4.550	7.175 ± 7.136	7.581 ± 0.139	$\mathbf{60.084 \pm 8.728}$	0.454 ± 0.065	1.329 ± 0.148	1.205 ± 1.099
Ours	3.385 ± 1.921	4.055 ± 2.584	$\mathbf{7.682 \pm 0.272}$	59.586 ± 12.197	$\mathbf{0.665 \pm 0.055}$	$\mathbf{1.855 \pm 0.257}$	$\mathbf{2.337 \pm 1.255}$

Table 6: Quantitative comparison on M^{3}SVD under the unified sequence-level protocol. All methods are retrained from official implementations on the UAV-VIVID source split and tested without target-domain fine-tuning on the identical 640×480 test list using the same metrics. Results are reported as mean $\pm$ standard deviation.

Methods [Pub./Year]	SEN↑	STD↑	SSIM↑	SCD↑	NQM↑
A2RNet [AAAI'25]	6.929 ± 0.528	46.814 ± 12.672	0.765 ± 0.078	0.543 ± 0.263	13.703 ± 2.055
DAFusion [IF'25]	6.576 ± 0.679	46.910 ± 15.294	0.743 ± 0.047	0.716 ± 0.207	12.076 ± 2.229
DCEvo [CVPR'25]	6.977 ± 0.539	46.376 ± 13.345	0.825 ± 0.073	0.342 ± 0.309	12.137 ± 1.795
MaeFuse [TIP'25]	7.101 ± 0.530	$\mathbf{54.774 \pm 15.033}$	0.793 ± 0.054	1.393 ± 0.142	9.712 ± 1.288
RIS-Fuse [ICCV'25]	6.932 ± 0.590	45.820 ± 12.338	0.842 ± 0.069	0.538 ± 0.237	12.961 ± 1.914
TemCoCo [ICCV'25]	6.658 ± 0.796	48.537 ± 13.450	0.744 ± 0.054	0.588 ± 0.275	9.611 ± 2.020
UniVF [NeurIPS'25]	7.026 ± 0.536	51.106 ± 14.329	0.823 ± 0.074	1.189 ± 0.227	12.389 ± 1.510
UP-Fusion [AAAI'26]	6.921 ± 0.497	46.172 ± 12.355	0.851 ± 0.075	0.466 ± 0.604	12.402 ± 1.994
AWM-Fuse [TIP'26]	6.637 ± 0.534	36.514 ± 10.048	0.843 ± 0.073	0.725 ± 0.384	10.862 ± 1.586
VideoFusion [CVPR'26]	6.455 ± 0.662	30.192 ± 9.867	0.631 ± 0.070	1.070 ± 0.366	5.202 ± 1.682
SDSFusion [TIP'25]	7.194 ± 0.221	52.306 ± 7.954	0.562 ± 0.123	0.567 ± 0.521	5.699 ± 3.235
Ours	$\mathbf{7.255 \pm 0.555}$	53.895 ± 12.395	$\mathbf{0.885 \pm 0.060}$	$\mathbf{1.455 \pm 0.366}$	$\mathbf{13.951 \pm 1.686}$

Table 7: Quantitative comparison on MSRS under the unified image-level protocol. All methods are retrained from official implementations on the FMB source split and tested without target-domain fine-tuning on the identical 640×480 test list using the same metrics. Results are reported as mean $\pm$ standard deviation.

Methods [Pub./Year]	SEN↑	STD↑	SSIM↑	SCD↑	NQM↑
A2RNet [AAAI'25]	7.010 ± 0.274	40.650 ± 5.371	0.588 ± 0.073	1.374 ± 0.131	3.699 ± 1.002
DAFusion [IF'25]	5.923 ± 0.358	26.609 ± 6.039	0.561 ± 0.085	1.082 ± 0.177	3.722 ± 1.640
DCEvo [CVPR'25]	7.096 ± 0.312	43.230 ± 6.859	0.661 ± 0.062	1.524 ± 0.177	3.390 ± 0.915
MaeFuse [TIP'25]	6.947 ± 0.268	40.278 ± 5.968	0.644 ± 0.066	1.630 ± 0.108	2.896 ± 0.899
RIS-Fuse [ICCV'25]	6.882 ± 0.352	38.613 ± 7.928	0.659 ± 0.066	1.375 ± 0.199	2.717 ± 1.188
TemCoCo [ICCV'25]	6.566 ± 0.390	35.995 ± 6.944	0.600 ± 0.064	1.178 ± 0.256	2.363 ± 1.364
UniVF [NeurIPS'25]	7.117 ± 0.308	44.458 ± 6.769	0.647 ± 0.063	1.679 ± 0.130	3.274 ± 0.904
UP-Fusion [AAAI'26]	6.963 ± 0.393	41.419 ± 7.329	0.676 ± 0.066	1.439 ± 0.228	3.499 ± 0.999
AWM-Fuse [TIP'26]	6.483 ± 0.324	29.169 ± 4.628	0.640 ± 0.081	1.025 ± 0.203	$\mathbf{3.903 \pm 1.212}$
VideoFusion [CVPR'26]	6.228 ± 0.438	24.185 ± 8.666	0.646 ± 0.093	0.636 ± 0.449	2.163 ± 1.219
SDSFusion [TIP'25]	$\mathbf{7.293 \pm 0.175}$	$\mathbf{47.502 \pm 7.711}$	0.328 ± 0.080	1.149 ± 0.146	0.200 ± 1.542
Ours	7.249 ± 0.368	45.693 ± 6.969	$\mathbf{0.687 \pm 0.079}$	$\mathbf{1.753 \pm 0.255}$	3.724 ± 1.256

Table 8: Quantitative comparison on LLVIP under the unified image-level protocol. All methods are retrained from official implementations on the FMB source split and tested without target-domain fine-tuning on the identical 640×480 test list using the same metrics. Results are reported as mean $\pm$ standard deviation.

Additional Quantitative and Qualitative Comparisons

Tables 9 and 10 report additional quantitative results on the FMB and M3FD datasets. On FMB, RT-FuseNet achieves the best results on four of the five directional metrics, namely SEN, STD, SSIM, and SCD, while obtaining the second-best NQM. On M3FD, it ranks first in SEN and SCD and second in STD and SSIM. Its NQM of 2.607 is also close to the best value of 2.661. These results indicate that our method provides a favorable balance between information content, image contrast, structural preservation, and cross-modal complementarity. The qualitative comparisons in Fig. 5 further support these observations. In the M3FD example, our result maintains conspicuous thermal responses for pedestrians while preserving the building contours, pavement structures, and overall scene appearance. In contrast, several competing methods either weaken the thermal targets or introduce over-smoothed and locally distorted appearances. In the FMB example, our method retains clearer traffic-sign boundaries and numerals, as well as a more distinguishable driver and vehicle-cabin structure. Meanwhile, the background textures and scene contrast are better preserved without severe overexposure. Overall, the visual results demonstrate that our method can effectively integrate infrared saliency with visible structural details across different scenes.

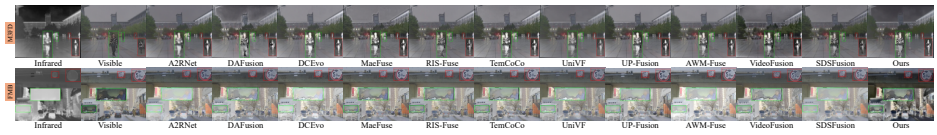


Figure 5: Qualitative comparison on the M3FD and FMB datasets. The infrared and visible inputs are followed by the fusion results of the competing methods and our method. The colored boxes highlight representative local regions for comparing thermal-target preservation, structural details, and local contrast.

Methods [Pub./Year]	SEN↑	STD↑	SSIM↑	SCD↑	NQM↑
A2RNet [AAAI'25]	6.598 ± 0.401	35.091 ± 7.963	0.728 ± 0.080	1.435 ± 0.225	1.439 ± 1.326
DAFusion [IF'25]	6.652 ± 0.446	31.715 ± 8.537	0.743 ± 0.085	1.527 ± 0.373	2.164 ± 1.146
DCEvo [CVPR'25]	6.849 ± 0.345	40.279 ± 9.177	0.722 ± 0.082	1.553 ± 0.121	1.084 ± 1.209
MaeFuse [TIP'25]	6.759 ± 0.343	34.771 ± 8.644	0.728 ± 0.080	1.519 ± 0.158	1.996 ± 1.474
RIS-Fuse [ICCV'25]	6.612 ± 0.373	32.858 ± 8.276	0.729 ± 0.078	1.376 ± 0.206	$\mathbf{2.876 \pm 2.001}$
TemCoCo [ICCV'25]	6.900 ± 0.279	39.027 ± 7.579	0.731 ± 0.084	1.610 ± 0.130	1.776 ± 1.574
UniVF [NeurIPS'25]	6.575 ± 0.303	29.958 ± 5.435	$\underline{0.761 \pm 0.078}$	1.535 ± 0.203	2.301 ± 1.139
UP-Fusion [AAAI'26]	6.901 ± 0.326	41.012 ± 9.928	0.708 ± 0.079	1.453 ± 0.139	0.778 ± 1.222
AWM-Fuse [TIP'26]	$\underline{6.923 \pm 0.292}$	41.906 ± 9.384	0.725 ± 0.085	$\underline{1.618 \pm 0.124}$	0.944 ± 1.157
VideoFusion [CVPR'26]	6.869 ± 0.388	$\underline{44.987 \pm 8.453}$	0.638 ± 0.051	1.587 ± 0.193	0.637 ± 0.969
SDSFusion [TIP'25]	6.524 ± 0.346	31.760 ± 7.906	0.736 ± 0.078	1.339 ± 0.217	2.133 ± 1.551
Ours	$\mathbf{7.014 \pm 0.346}$	$\mathbf{45.539 \pm 7.132}$	$\mathbf{0.764 \pm 0.080}$	$\mathbf{1.705 \pm 0.165}$	$\underline{2.316 \pm 0.974}$

Table 9: Quantitative comparison on FMB under the unified image-level protocol. All methods are retrained on the same FMB source training split and evaluated on the identical 640×480 FMB test list using the same metric implementations. Results are reported as mean $\pm$ standard deviation.

Methods [Pub./Year]	SEN↑	STD↑	SSIM↑	SCD↑	NQM↑
A2RNet [AAAI'25]	6.774 ± 0.366	34.566 ± 8.704	0.681 ± 0.095	1.428 ± 0.219	1.050 ± 1.304
DAFusion [IF'25]	6.847 ± 0.777	35.144 ± 9.767	0.680 ± 0.126	$\underline{1.616 \pm 0.397}$	1.386 ± 1.254
DCEvo [CVPR'25]	7.003 ± 0.414	41.298 ± 10.130	0.689 ± 0.083	1.516 ± 0.160	0.933 ± 1.222
MaeFuse [TIP'25]	6.956 ± 0.390	37.050 ± 8.943	0.693 ± 0.086	1.540 ± 0.178	1.618 ± 1.362
RIS-Fuse [ICCV'25]	6.820 ± 0.395	35.166 ± 8.014	0.692 ± 0.083	1.353 ± 0.248	$\mathbf{2.661 \pm 1.843}$
TemCoCo [ICCV'25]	7.055 ± 0.357	40.064 ± 8.388	0.691 ± 0.087	1.587 ± 0.157	1.646 ± 1.461
UniVF [NeurIPS'25]	6.635 ± 0.405	29.580 ± 7.344	$\mathbf{0.726 \pm 0.086}$	1.577 ± 0.257	1.947 ± 1.090
UP-Fusion [AAAI'26]	7.049 ± 0.349	42.520 ± 9.602	0.671 ± 0.082	1.451 ± 0.183	0.526 ± 1.164
AWM-Fuse [TIP'26]	6.804 ± 0.409	35.055 ± 8.219	0.714 ± 0.085	1.499 ± 0.205	1.998 ± 1.351
VideoFusion [CVPR'26]	$\underline{7.080 \pm 0.346}$	$\mathbf{46.356 \pm 9.117}$	0.599 ± 0.068	1.585 ± 0.230	0.394 ± 1.061
SDSFusion [TIP'25]	6.658 ± 0.528	31.718 ± 9.156	0.684 ± 0.092	1.570 ± 0.270	$\underline{2.647 \pm 1.346}$
Ours	$\mathbf{7.121 \pm 0.257}$	$\underline{43.175 \pm 7.062}$	$\underline{0.722 \pm 0.072}$	$\mathbf{1.661 \pm 0.181}$	2.607 ± 1.864

Table 10: Cross-dataset quantitative comparison on M3FD. All methods are retrained on the same FMB source training split and directly evaluated on the identical 640×480 M3FD test list without target-domain fine-tuning, using the same metric implementations. Results are reported as mean $\pm$ standard deviation.

Downstream Object Detection

Evaluation protocol and quantitative results. We evaluate pedestrian detection on LLVIP under a shared YOLOv8s protocol. RT-FuseNet achieves the highest precision of 0.8598 and the best $AP_{50:95}$ of 0.6425, while obtaining the second-best AP_{50} of 0.8461 and a competitive recall of 0.7591.

Qualitative results. Figure 6 shows qualitative downstream pedestrian detection on LLVIP. The fused images support more complete pedestrian localization in low-light and shadowed scenes, particularly for small or partially occluded targets, while retaining visible structural cues.

Method	P↑	R↑	AP_{50} ↑	$AP_{50:95}$ ↑	Method	P↑	R↑	AP_{50} ↑	$AP_{50:95}$ ↑
IR	0.8048	0.7458	**0.9365**	0.6224	TemCoCo	0.8331	0.7483	0.8362	0.5839
VI	0.8124	0.7124	0.8450	0.4703	SDSFusion	0.8364	0.7222	0.8327	0.5912
A2RNet	0.8067	0.7162	0.8450	0.6219	UniVF	0.8235	0.7466	0.8263	0.5831
DAFusion	0.8497	0.7075	0.8071	0.5646	UP-Fusion	0.8173	0.7587	0.8358	0.5976
DCEvo	0.8425	0.7603	0.8250	0.5934	AWM-Fuse	0.8408	0.7447	0.8324	0.6032
MaeFuse	0.8381	**0.7612**	0.8447	0.6030	VideoFusion	0.8371	0.7315	0.8361	0.6167
RIS-Fuse	0.8447	0.7385	0.8410	0.6010	**Ours**	**0.8598**	0.7591	0.8461	**0.6425**

Table 11: Pedestrian detection on LLVIP. Best and second-best results are shown in bold and underlined, respectively.

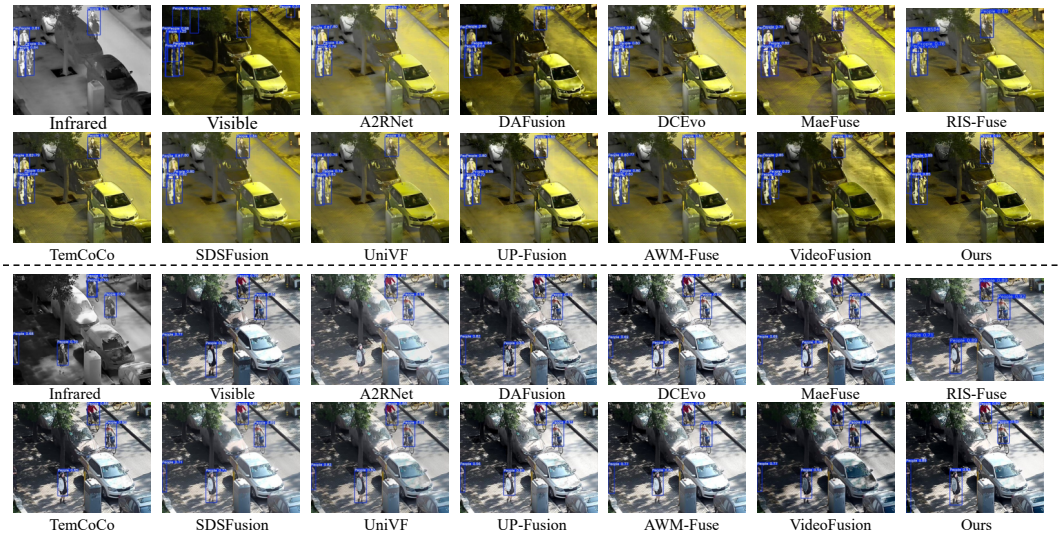


Figure 6: Qualitative downstream pedestrian detection on LLVIP.

Downstream Semantic Segmentation

Evaluation protocol and quantitative results. Following the SegMiF protocol, we evaluate downstream semantic segmentation on MSRS using the same segmentation architecture and training settings for all fusion methods. For each method, the segmenter is trained on the corresponding fused MSRS training set and evaluated on the same test set. Table 12 reports the class-wise accuracy and intersection over union (IoU), as well as their mean values. RT-FuseNet achieves the highest mAcc of 89.80 and the second-highest mIoU of 77.45, only 0.03 below the best result. These results demonstrate that its fused images retain complementary infrared and visible cues that are beneficial for downstream semantic segmentation.

Qualitative results. Figure 7 presents semantic segmentation results on FMB. Competing methods may produce fragmented regions, category confusion, or inaccurate boundaries under low illumination and clutter. RT-FuseNet yields more complete regions and clearer boundaries for pedestrians, vehicles, roads, sidewalks, and vegetation, illustrating its qualitative behavior for downstream semantic perception.

Methods [Pub./Year]	Car		Person		Bike		Curve		Car_stop		Guardrail		Color_cone		Bump		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
Visible	94.99	89.57	79.58	61.76	85.96	71.05	80.08	61.32	88.97	71.69	92.17	82.51	77.83	62.17	87.68	73.22	87.36	74.62
Infrared	94.75	86.93	83.08	71.24	84.82	69.30	79.86	56.76	83.86	61.71	83.65	60.07	80.96	57.83	88.82	70.39	86.51	70.27
A2RNet [AAAI'25]	94.77	89.50	84.51	72.07	85.65	70.90	81.08	63.39	87.00	75.35	91.94	83.89	77.91	62.36	88.19	76.42	87.80	76.93
DAFusion [IF'25]	95.30	90.29	86.06	74.93	86.08	71.85	82.70	64.01	89.55	74.98	92.57	83.34	79.70	63.04	85.53	73.95	88.52	77.22
DCEvo [CVPR'25]	95.25	90.03	85.04	74.82	85.96	71.31	81.71	62.82	88.91	75.61	91.57	81.26	85.58	66.03	90.14	76.93	89.26	**77.48**
MaeFuse [TIP'25]	95.50	90.63	85.93	74.51	85.92	**71.98**	82.35	62.34	89.56	72.79	91.34	82.92	79.48	63.12	89.29	77.35	88.72	77.13
RIS-Fuse [ICCV'25]	95.16	90.90	85.50	74.17	84.70	70.84	81.16	63.03	87.53	72.79	90.68	85.46	80.35	64.51	89.30	74.47	88.17	77.19
TemCoCo [ICCV'25]	95.46	89.49	84.60	73.46	86.52	70.74	**84.44**	59.36	87.52	72.28	91.38	84.33	84.44	59.38	**91.51**	69.20	89.43	75.18
SDSFusion [TIP'25]	94.86	88.79	87.23	68.15	85.39	68.31	82.06	57.26	85.59	72.97	92.66	**86.93**	78.66	61.34	86.76	70.05	88.01	74.68
UniVF [NeurIPS'25]	95.54	**91.19**	86.79	**75.47**	86.35	71.59	82.34	64.15	**90.02**	74.76	91.84	80.43	84.69	**67.04**	89.88	73.72	89.62	77.44
UP-Fusion [AAAI'26]	95.31	90.19	86.22	75.18	84.96	70.26	83.26	63.20	87.96	71.63	92.92	81.71	78.20	61.31	90.40	75.23	88.70	76.36
AWM-Fuse [TIP'26]	**95.84**	90.04	85.98	74.89	85.91	70.76	81.13	62.92	87.71	74.54	**93.37**	83.98	79.81	61.99	89.80	74.45	88.74	76.90
VideoFusion [CVPR'26]	94.96	89.32	84.58	70.61	**87.19**	68.97	79.17	54.51	85.70	70.49	92.39	78.25	80.80	60.56	88.48	72.40	88.02	73.71
Ours	**95.84**	90.92	**87.52**	75.22	86.98	71.86	84.17	**64.25**	89.87	**75.87**	92.99	86.86	**85.84**	66.50	90.88	**77.96**	**89.80**	77.45

Table 12: Segmentation results on the MSRS dataset. The best and second-best results are highlighted in bold and underlined, respectively.

Methods [Pub./Year]	Road		Building		Lamp		Vegetation		Sky		Person		Car		Truck		Bus		Motorcycle		Pole		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
Visible	86.85	84.02	88.93	77.79	77.17	25.29	88.60	81.97	95.88	92.71	76.24	42.93	82.22	74.72	**72.22**	16.83	73.70	17.50	65.72	38.83	67.80	46.27	75.93	52.13
Infrared	84.53	82.55	88.03	78.02	76.67	11.19	87.54	78.98	95.38	92.69	86.46	57.14	81.63	73.01	54.91	12.24	58.68	18.11	40.95	26.01	58.30	28.64	71.03	47.52
ATGan-Fusion [IF'23]	84.56	81.70	85.66	76.48	65.86	28.64	91.55	80.96	93.99	91.39	79.62	52.87	80.53	72.15	38.90	7.16	67.23	10.96	59.71	34.60	66.07	43.86	66.93	46.97
IF-FILM [ICML'24]	87.74	84.97	87.53	81.14	78.45	35.25	93.41	84.59	95.96	93.80	85.59	60.41	81.69	74.94	68.78	24.53	66.24	14.58	54.02	34.97	73.11	44.75	75.62	54.45
Text-IF [CVPR'24]	87.18	84.72	86.34	79.92	79.37	32.86	92.87	84.69	96.43	93.74	82.77	59.37	82.17	75.27	51.34	18.92	72.83	20.45	52.65	33.13	72.63	46.36	75.75	54.23
PromptFusion [ECCV'24]	87.38	85.00	87.37	79.66	77.24	25.49	92.19	84.37	96.06	93.26	84.68	59.35	81.58	74.74	67.84	**33.83**	78.24	14.38	65.37	43.40	71.72	44.52	**77.43**	55.02
Text-Fusion [IF'25]	86.94	84.61	86.72	79.49	76.50	30.73	91.88	83.92	96.19	92.94	84.15	**64.12**	**83.04**	**76.04**	40.66	17.57	72.88	14.56	69.44	43.24	70.63	45.88	70.60	50.93
DCEvo [CVPR'25]	87.34	85.19	87.93	79.85	77.61	**36.12**	91.61	84.67	96.09	93.50	85.79	58.57	82.52	74.99	49.33	14.09	77.18	17.04	62.26	38.47	73.10	45.48	76.36	54.42
CrossFuse [TCSVT'25]	87.91	85.33	87.51	79.95	78.87	32.66	90.85	84.51	96.67	93.85	84.96	58.98	82.59	75.27	41.24	12.51	75.01	13.18	50.10	35.75	72.85	46.87	74.97	53.47
A2RNet [AAAI'25]	85.40	83.03	85.54	78.53	71.48	19.23	92.60	83.65	95.93	93.01	86.64	54.27	80.66	72.44	33.48	6.95	60.07	9.29	56.44	31.16	71.60	44.43	71.95	49.84
CSDFusion [KBS'26]	88.11	84.38	86.89	80.85	76.00	35.27	92.18	84.64	96.86	93.89	84.24	62.84	82.16	75.71	58.14	15.20	**83.12**	19.17	62.74	44.39	71.67	47.20	77.14	55.50
CPFusion [IVC'26]	87.97	85.28	87.95	79.03	79.13	29.75	91.90	84.66	95.69	93.20	86.46	56.88	80.61	74.48	63.57	18.12	80.41	16.45	53.13	33.69	70.82	46.63	71.35	49.92
ZVIR [PR'26]	86.92	84.71	88.94	78.84	76.27	26.65	90.97	84.04	95.69	93.33	83.58	56.57	79.60	72.95	57.34	15.30	80.14	16.06	59.36	40.19	72.59	45.65	75.76	53.06
Ours	**88.98**	**85.97**	**89.12**	**82.92**	**82.72**	36.04	**93.47**	**85.98**	**96.97**	**94.51**	**86.82**	63.84	82.62	75.94	69.58	30.72	81.39	**21.84**	**69.54**	**44.57**	**73.93**	**48.79**	77.34	**56.07**

Table 13: Segmentation results on the FMB dataset. The best and second-best results are highlighted in bold and underlined, respectively.

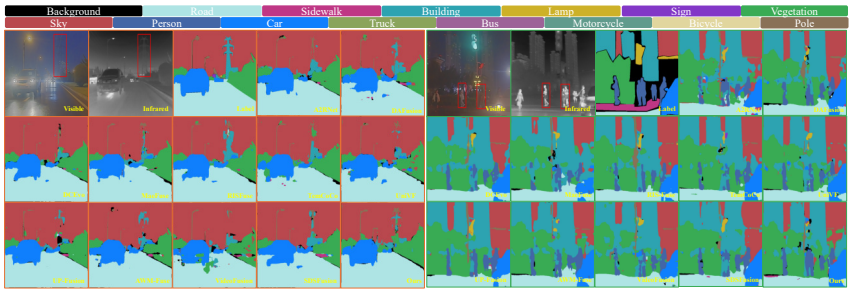


Figure 7: Qualitative downstream semantic segmentation results on the FMB dataset.

Computational efficiency. Table 14 reports computational statistics at an input resolution of 640×480. All inference times are remeasured on a single NVIDIA L20 GPU with a batch size of 1 using the same precision setting. After warm-up iterations, latency is averaged over repeated runs. Data loading, file I/O, and offline preprocessing are excluded.

Structured descriptions and their BLIP-2 embeddings are generated offline and cached before fusion inference. We therefore report fusion-network-only latency in Table 14. RT-FuseNet contains 6.25 M parameters, requires 158.97 G FLOPs, and has an average fusion-network latency of 1332.20 ms per IR–VI pair.

Method [Pub./Year]	Time (ms)↓	FLOPs (G)↓	Params (M)↓	Method [Pub./Year]	Time (ms)↓	FLOPs (G)↓	Params (M)↓
A2RNet [AAAI'25]	3154.70	164.82	3.57	SDSFusion [TIP'25]	97.40	89.79	0.68
DAFusion [IF'25]	8220.00	62.07	7.92	UniVF [NeurIPS'25]	105.90	280.52	3.54
DCEvo [CVPR'25]	748.41	2637.23	2.00	UP-Fusion [AAAI'26]	142.46	231.60	0.49
MaeFuse [TIP'25]	300.00	160.42	90.46	AWM-Fuse [TIP'26]	505.18	530.53	7.44
RIS-Fuse [ICCV'25]	13.80	3624.00	11.80	VideoFusion [CVPR'26]	2.32	0.16	0.01
TemCoCo [ICCV'25]	22.60	547.74	1.19	**Ours**	1332.20	158.97	6.25

Table 14: Computational-efficiency comparison of representative IR–VI fusion methods at an input resolution of 640×480, reported in terms of inference time, FLOPs, and parameter count.

2 Additional Ablation Results

Effect of Training-Time PGD Augmentation

Table 15 compares RT-FuseNet trained with and without joint dual-modal PGD augmentation. On clean validation inputs, PGD augmentation reduces BiSWE from 4.593 to 4.443 and flowD from 7.453 to 7.296, while increasing SSIM from 0.544 to 0.595 and NQM from 4.102 to 4.352. These results suggest that training-time perturbation acts as an effective regularizer for clean-input fusion without changing the inference procedure. Figure 8 provides corresponding visual comparisons of the prompt and training variants.

Training	BiSWE↓	flowD↓	SSIM↑	NQM↑
w/o PGD aug.	4.593 ± 2.674	7.453 ± 3.264	0.544 ± 0.085	4.102 ± 1.811
w/ PGD aug.	$\mathbf{4.443 \pm 2.836}$	$\mathbf{7.296 \pm 3.492}$	$\mathbf{0.595 \pm 0.054}$	$\mathbf{4.352 \pm 1.798}$

Table 15: Clean-input effect of PGD training augmentation on UAV-VIVID (mean $\pm$ standard deviation).

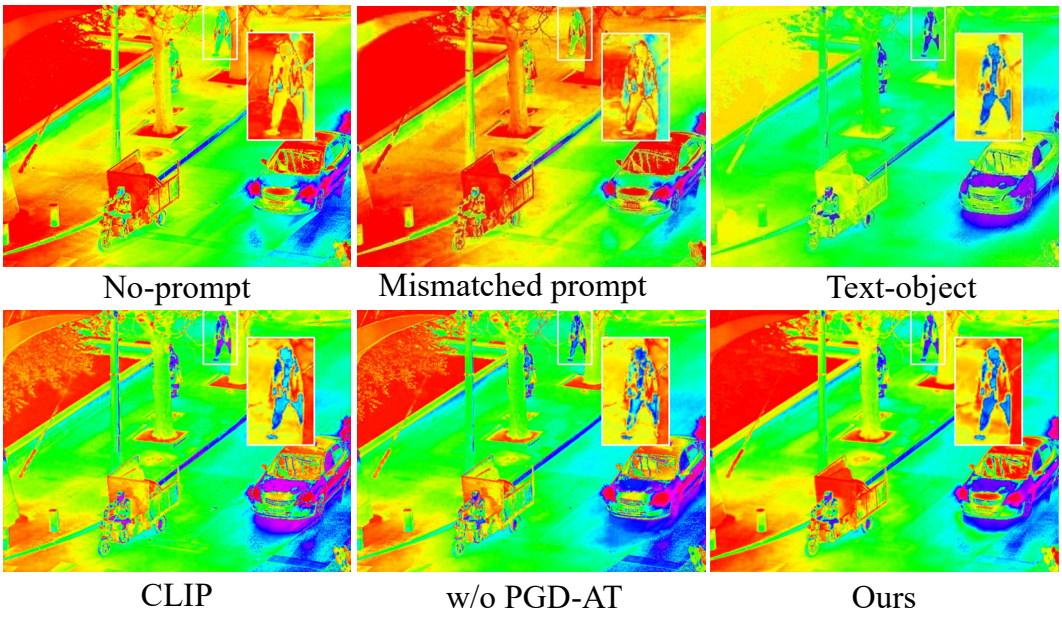


Figure 8: Clean-input qualitative ablation on UAV-VIVID across prompt, encoder, and PGD-training variants. "Text-object" denotes the object-only prompt, while "w/o PGD-AT" denotes the removal of training-time PGD augmentation. No adversarially perturbed input is shown.

ARB AND TAGM ABLATION WITH VARIABILITY

Effectiveness of ARB and TAGM. Table 16 shows that both ARB and TAGM improve the plain backbone. Their combination, M4, achieves the best BiSWE, flowD, SEN, and SSIM, improving BiSWE/flowD/SSIM from 4.862/7.652/0.535 to 4.443/7.296/0.595. Although M3 obtains the highest STD, M4 provides the best overall balance across the reported metrics, supporting the complementary roles of cross-modal balancing and text-conditioned modulation.

Model	TAGM	ARB	BiSWE ↓	flowD ↓	SEN ↑	STD ↑	SSIM ↑
M1	✗	✗	4.862 ± 2.541	7.652 ± 2.633	6.850 ± 0.610	45.701 ± 10.266	0.535 ± 0.080
M2	✗	✔	4.769 ± 2.635	7.563 ± 3.541	7.215 ± 0.693	55.678 ± 11.595	0.571 ± 0.071
M3	✔	✗	4.815 ± 2.637	7.365 ± 3.569	7.150 ± 0.697	**58.501 ± 11.673**	0.582 ± 0.069
M4	✔	✔	**4.443 ± 2.836**	**7.296 ± 3.492**	**7.352 ± 0.673**	56.249 ± 20.153	**0.595 ± 0.054**

Table 16: Clean-input ARB/TAGM ablation on UAV-VIVID under identical settings (mean ± standard deviation).

PROMPT AND TEXT-ENCODER ABLATION WITH VARIABILITY

Table 17 provides the complete mean-and-standard-deviation results for the prompt and text-encoder ablation reported in the main paper. All variants use the same training-time PGD-augmentation configuration and are evaluated on clean inputs. The best and second-best directional metrics are shown in bold and underlined, respectively. BiSWE and flowD describe the output-sequence behavior of independently fused frames and do not imply temporal context.

Prompt	BiSWE ↓	flowD ↓	SEN ↑	STD ↑	SSIM ↑
No-prompt	5.495 ± 2.364	8.716 ± 3.169	6.807 ± 0.644	47.485 ± 20.557	0.403 ± 0.066
CLIP	4.835 ± 3.641	7.516 ± 2.785	7.187 ± 0.618	49.536 ± 19.793	0.566 ± 0.054
Text-object prompt	4.546 ± 2.124	7.331 ± 3.164	7.063 ± 0.605	55.394 ± 19.766	0.569 ± 0.059
Mismatched prompt	4.941 ± 2.694	8.249 ± 3.367	6.928 ± 0.635	48.824 ± 21.694	0.485 ± 0.098
Scene + object prompt (Ours)	**4.443 ± 2.836**	**7.296 ± 3.492**	**7.352 ± 0.673**	**56.249 ± 20.153**	**0.595 ± 0.054**

Table 17: Ablation study of different prompt and text-encoder configurations on the fixed scene-disjoint 20-scene validation split of the registration-processed lower-altitude UAV-VIVID benchmark. All variants are trained on the same 109-scene training split under identical settings. Results are reported as mean ± standard deviation.

CONTROLLED TEMPORAL ANALYSIS

We define three diagnostic variants: RT-FuseNet-1F uses only the center frame t; RT-FuseNet-3F uses frames $t-1, t, t+1$; and RT-FuseNet-5F uses frames $t-2, t-1, t, t+1, t+2$. The 1F variant applies the original RT-FuseNet bottleneck representation directly. For 3F and 5F, the same RT-FuseNet encoder is shared across all frames in the temporal window. At the encoder bottleneck, a lightweight temporal-attention adapter uses the center feature as query and the window features as keys and values. Attention is computed across time at each bottleneck spatial location:

$$\begin{aligned} Q_t &= W_Q F_t, \qquad K_j = W_K F_{t+j}, \qquad V_j = W_V F_{t+j}, \\ a_j(p) &= \mathrm{softmax}_j \left(\frac{Q_t(p)^\top K_j(p)}{\sqrt{d}} \right), \\ \widetilde{F}_t &= F_t + W_O \left(\sum_j a_j \odot V_j \right). \end{aligned} \tag{15}$$

The adapter output is residually added to the center feature and is passed to TAGM. Neighboring-frame features do not enter decoder skip connections: every skip connection uses only the corresponding center-frame encoder feature. All variants use the center-frame text description and

predict only the center fused frame. No optical-flow alignment, Mamba, ConvLSTM, deformable alignment, or other temporal architecture is added.

For this controlled experiment, PGD is disabled for all three variants. They use the same scene-disjoint training split, training objective, optimizer, learning rate, epoch count, effective batch size, and text conditioning. Temporal windows are restricted to frames within one scene. Evaluation uses the exact same center-frame list for all variants: the valid center frames for the 5F window, for which all two preceding and two following frames exist in the same scene. This common list also defines the center frames evaluated by 1F and 3F. The variants are diagnostic baselines for testing the value of temporal context in UAV-VIVID, not new primary methods.

Variant	Frames	Params (M)	BiSWE ↓	flowD ↓	SEN ↑	STD ↑	SSIM ↑	SCD ↑	NQM ↑
RT-FuseNet-1F	1	6.25	5.287 ± 2.758	7.277 ± 3.527	7.719 ± 0.258	60.597 ± 9.942	0.678 ± 0.049	1.655 ± 0.258	4.579 ± 1.643
RT-FuseNet-3F	3	6.25	5.146 ± 3.874	7.125 ± 3.642	7.652 ± 0.342	60.425 ± 8.526	0.526 ± 0.159	1.625 ± 0.346	4.512 ± 1.566
RT-FuseNet-5F	5	6.25	5.194 ± 3.913	7.107 ± 3.931	7.693 ± 0.493	60.412 ± 0.136	0.515 ± 0.123	1.606 ± 0.369	4.499 ± 1.789

Table A1: Controlled evaluation of temporal context on UAV-VIVID.

Effect of Temporal Context. On the common valid center-frame test list, RT-FuseNet-1F, -3F, and -5F obtain BiSWE values of 5.287, 5.146, and 5.194, respectively, while their flowD scores decrease from 7.277 to 7.125 and 7.107. These results indicate that incorporating neighboring frames can improve sequence-level consistency, although extending the temporal window from three to five frames provides only marginal additional benefit and slightly degrades BiSWE. In contrast, the frame-wise fusion metrics do not exhibit the same trend: the single-frame baseline retains the highest SEN, SSIM, SCD, and NQM, while both temporal variants show lower spatial-quality scores, with the reduction being particularly evident for SSIM. Overall, the controlled comparison reveals a trade-off between improved temporal consistency and preservation of frame-wise fusion quality, suggesting that directly aggregating neighboring-frame information is insufficient to simultaneously optimize both objectives (Table A1).

3 Ethical Collection, Privacy, and Release Governance

Collection authorization. All UAV operations were conducted by certified personnel. Data acquisition was limited to locations where permission had been granted by the relevant institutional or site authorities. The collection was observational and non-interventional, and the UAV was not used to deliberately track or identify specific individuals. No audio, names, contact information, or other directly provided personal identifiers were collected.

Privacy filtering. Before public release, visible frames are processed using automatic face and license-plate detection followed by manual inspection of high-risk and high-resolution scenes. Detected identity-bearing regions are blurred. Exact GPS coordinates, device identifiers, and other metadata that are not required for benchmark evaluation are removed or coarsened. Infrared data are not treated as inherently anonymous, because temporal trajectories and contextual information may still create re-identification risks.

Release governance. The public release contains only privacy-filtered imagery and task-relevant metadata. Access to any unredacted or fine-grained source material, if retained, is restricted and requires an institutional application and a signed data-use agreement. The agreement prohibits re-identification, face recognition, individual surveillance, redistribution, and attempts to recover removed location metadata.

4 Synchronization and Registration Accuracy

This section quantifies the temporal pairing constraints and the geometric calibration accuracy of UAV-VIVID. We distinguish temporal matching residuals, intrinsic-calibration reprojection errors, and residual scene-level misalignment, as these measurements characterize different stages of the processing pipeline.

TEMPORAL SYNCHRONIZATION ACCURACY

The infrared and visible streams were recorded at a nominal frame rate of 30 fps, corresponding to a frame period of approximately 33.33 ms. Before temporal pairing, periodic near-duplicate infrared frames were detected and removed. For each manually anchored scene with reliable timestamps, a manually verified pair of infrared and visible frames was used as a temporal anchor. The timestamp difference between this anchor pair was applied as a constant offset to the infrared timestamp sequence.

Pairing mode	**Rate**	**Maximum residual**	**Frame equiv.**
Manual anchor + nearest timestamp	30 fps	25 ms	0.75

Table 18: Temporal pairing constraint for manually anchored sequences. The frame equivalent assumes 30 fps.

After anchor correction, infrared frames were used as the reference stream and matched one-to-one to the nearest unused visible frames. For the i-th retained frame pair, the signed and absolute timestamp residuals are defined as

$$d_i = t_i^{\mathrm{VI}} - t_i^{\mathrm{IR}}, \qquad \Delta t_i = |d_i|, \tag{16}$$

where t_i^{IR} and t_i^{VI} denote the offset-corrected matching timestamps.

A maximum admissible residual of 25 ms was used during final matching. Consequently, every retained frame pair in the manually anchored fusion sequences satisfies

$$\Delta t_i \leq 25 \text{ ms}, \tag{17}$$

which corresponds to at most 0.75 frame periods at 30 fps. Frames for which no unused cross-modal counterpart satisfied this bound were discarded rather than duplicated or interpolated. The synchronization procedure therefore preserves the original decoded frames and does not perform temporal resampling.

The reported 25-ms value is an upper bound on the residual between offset-corrected recorded timestamps. It should not be interpreted as shutter-level hardware synchronization, because the two sensors were not controlled by a shared external trigger. Residual appearance differences may also arise from rolling or non-simultaneous exposure, rapid object motion, and independently operating sensor pipelines.

GEOMETRIC CALIBRATION ACCURACY

Sensor	**Resolution**	**Valid views**	**RMS (px)**	**Diag. (%)**
Infrared	1280×1024	24	1.2868	0.0785
Visible	3840×2160	27	1.2128	0.0275

Table 19: Intrinsic-calibration errors at native sensor resolutions.

More than 5,000 raw checkerboard frames were recorded during the calibration procedure. After corner-detection and pose-diversity filtering, 24 infrared views and 27 visible views with valid corner observations were retained for intrinsic calibration. The checkerboard contains a 9×10 inner-corner grid with a physical square size of 20 mm.

Calibration accuracy is measured using the root mean squared reprojection error

$$e_{\mathrm{cal}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}_i - \widehat{\mathbf{x}}_i\|_2^2}, \tag{18}$$

where $\mathbf{x}_i$ denotes a detected checkerboard corner and $\widehat{\mathbf{x}}_i$ is its projection obtained using the estimated camera parameters.

The infrared camera obtains an RMS reprojection error of 1.2868 pixels at its native resolution of 1280×1024, corresponding to 0.0785% of the image diagonal. The visible camera obtains an RMS reprojection error of 1.2128 pixels at its native resolution of 3840×2160, corresponding to 0.0275% of the image diagonal. These values are small relative to the native image dimensions, although they do not imply perfect scene-level cross-modal alignment.

Residual Scene-Level Misalignment

For the registration-processed lower-altitude subset, visible frames are warped onto the infrared image plane using homographies estimated from manually verified cross-modal correspondences on representative keyframes. The resulting registered resolution is 1280×1024.

Intrinsic reprojection error and final scene-level registration error describe different quantities. The former measures how accurately the camera model reproduces checkerboard observations, whereas the latter is also affected by scene depth variation, parallax, moving objects, thermal diffusion, and the approximation of a non-planar scene using a single homography. Accordingly, we refer to this subset as *registration-processed* rather than perfectly pixel-aligned.

Dynamic objects, vegetation, water surfaces, reflective regions, diffuse thermal boundaries, and strong depth discontinuities may retain local displacement after registration. These residual discrepancies are intentionally preserved as part of the practical UAV acquisition setting. The 14,899 unregistered lower-altitude pairs are released separately and are not used in the fusion experiments reported in the paper.