



# ControlFusion: A Controllable Image Fusion Network with Language-Vision Degradation Prompts

Linfeng Tang<sup>1\*</sup>, Yeda Wang<sup>1\*</sup>, Zhanchuan Cai<sup>2</sup>, Junjun Jiang<sup>3</sup>, Jiayi Ma<sup>1†</sup>

<sup>1</sup>Wuhan University  <sup>2</sup>Macau University of Science and Technology  <sup>3</sup>Harbin Institute of Technology 

\*Equal Contribution    †Corresponding Author

 View on Github: <https://github.com/Linfeng-Tang/ControlFusion>

This work was supported by National Natural Science Foundation of China (No. 62276192).

## □ Introduction & Motivation

**Goal:** To synergize the **complementary strengths** of both modalities: integrating the **rich textural details** of **visible** imagery with the **salient thermal targets** of **infrared** data.



Imaging Wavelength



Imaging Principle



Advantages



Limitations



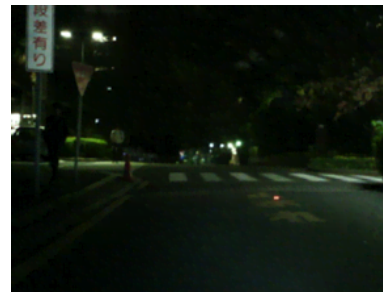
Examples

### Visible Modality

380-780 nm

Reflection-based Imaging

- ✓ Clear texture and appearance
- Severe information loss under poor lighting conditions or in the presence of camouflage



### Infrared Modality

8-14  $\mu\text{m}$

Thermal Radiation Imaging

- ✓ Highlights thermal targets (e.g., humans, vehicles)
- High noise levels, blurred details, and poor visual quality (lack of texture)





## □ Introduction & Motivation

### BROAD APPLICATION SCOPE



Safety



Reliability

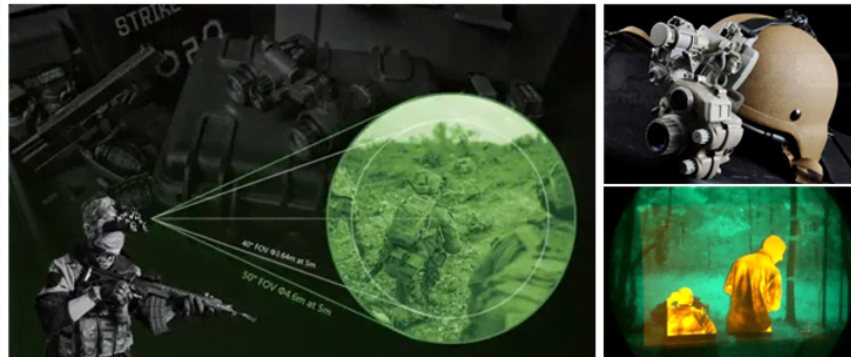


24-Hour Availability



Targeting complex environments, not just lab benchmarks

### Military Detection



DTG-18N Military Panoramic Night Vision Goggles

### Urban Security



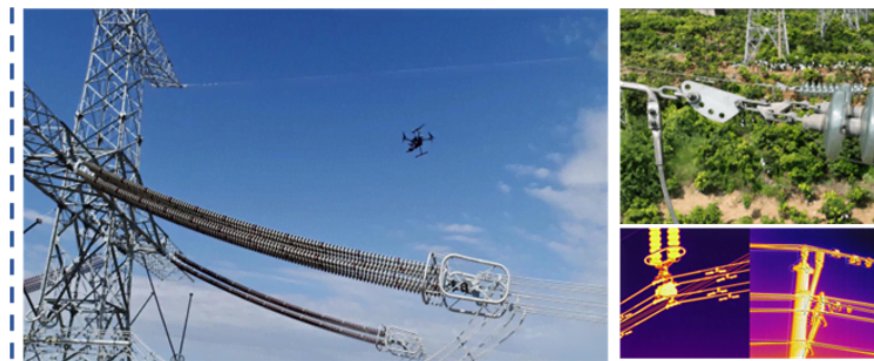
PTV-R Intelligent Security Surveillance System

### Autonomous Driving



QuadSight Night Driving Assistance System

### Intelligent Industry



UAV Multimodal Industrial Inspection Platform

## □ Introduction & Motivation

### High quality scenarios



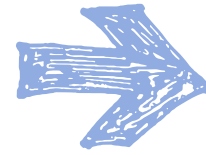
Sensor



Weather



Illumination



### Complex interference scenarios



### Fusion results in degradation scenarios



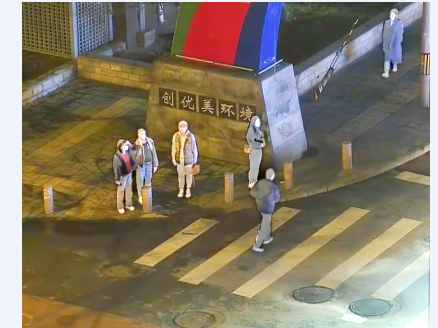
(a) Visible



(b) Infrared



(c) EMMA [CVPR 24']



(d) Ours

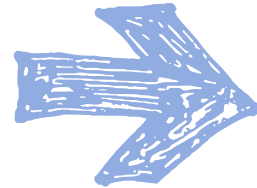
Vulnerability of SOTA methods to **degradation scenarios**



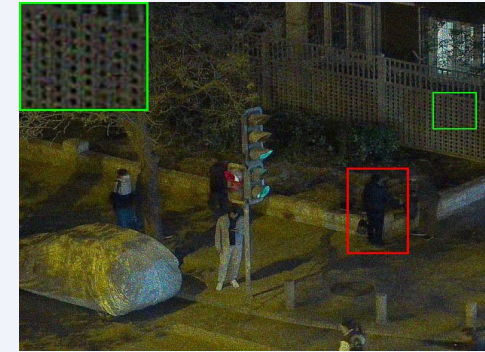
## □ Introduction & Motivation

### Degradation entanglement

- IR (Low contrast)
- IR (Noise)
- IR (Stripe noise)
- VI (Blur)
- VI (Haze)
- VI (Haze & Low)
- VI (Low light)
- VI (Noise)
- VI (Noise & Low)
- VI (Overexposure)
- VI (Rain)
- VI (Rain & Haze)



### Fusion results under composite degradation



(a) Visible



(b) Infrared



(c) Text-IF [CVPR 24']

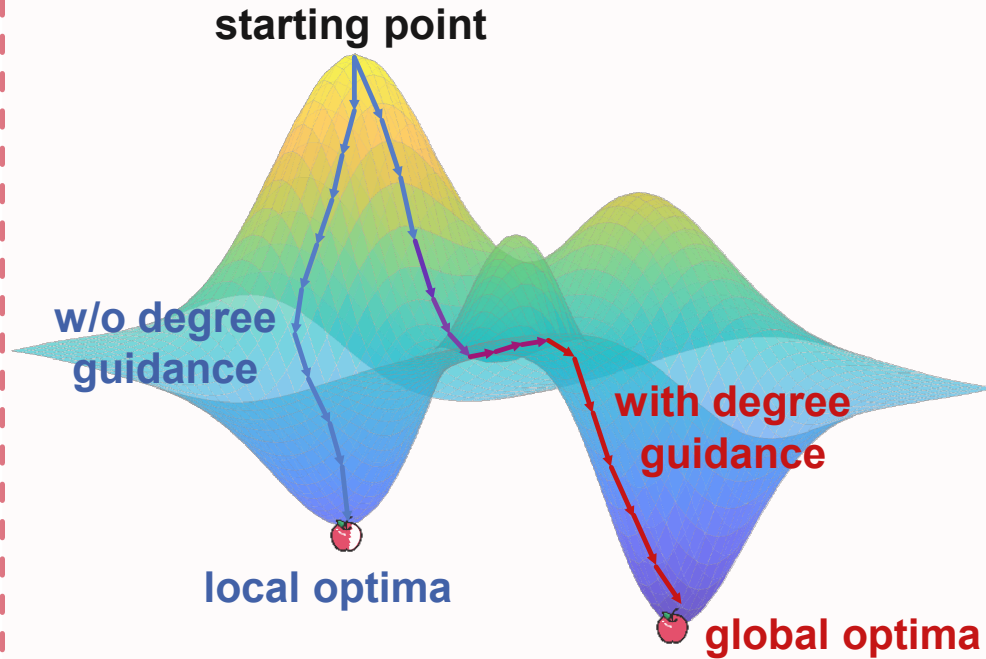


(d) Ours

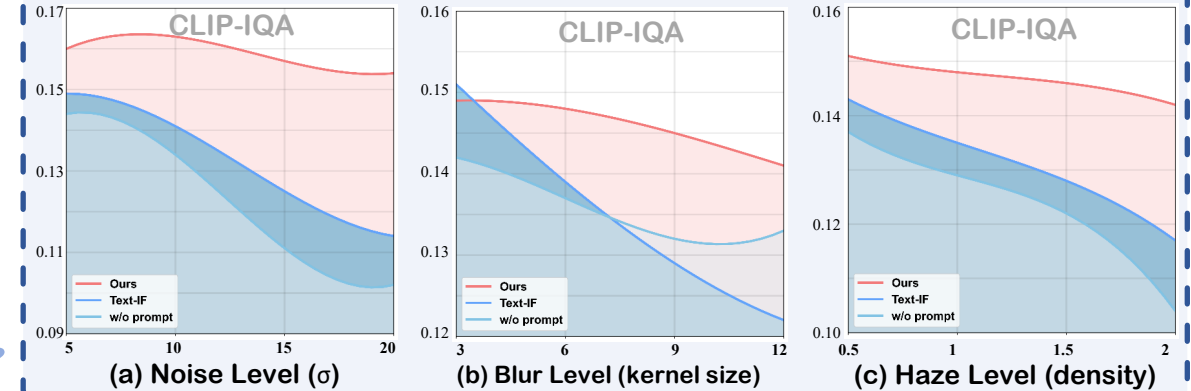
**Degradation entanglement** causes failure in degradation-aware models

## □ Introduction & Motivation

### Schematic diagram of gradient solution



### Degree sensitivity analysis

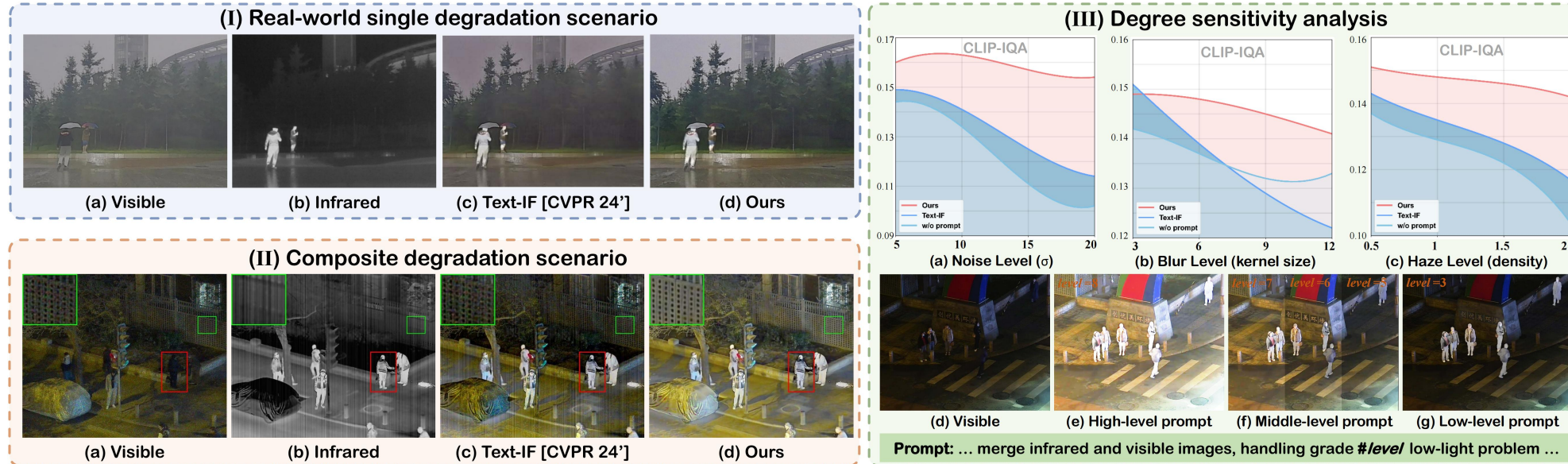


Prompt: ... merge infrared and visible images, handling grade **#/level/** low-light problem ...

Lack of **degree modeling** leads to suboptimal results



## □ Introduction & Motivation



Existing restoration-fusion methods overlooking the **domain gap** between **simulated** data and **realistic** images, which hampers their generalizability in practical scenarios

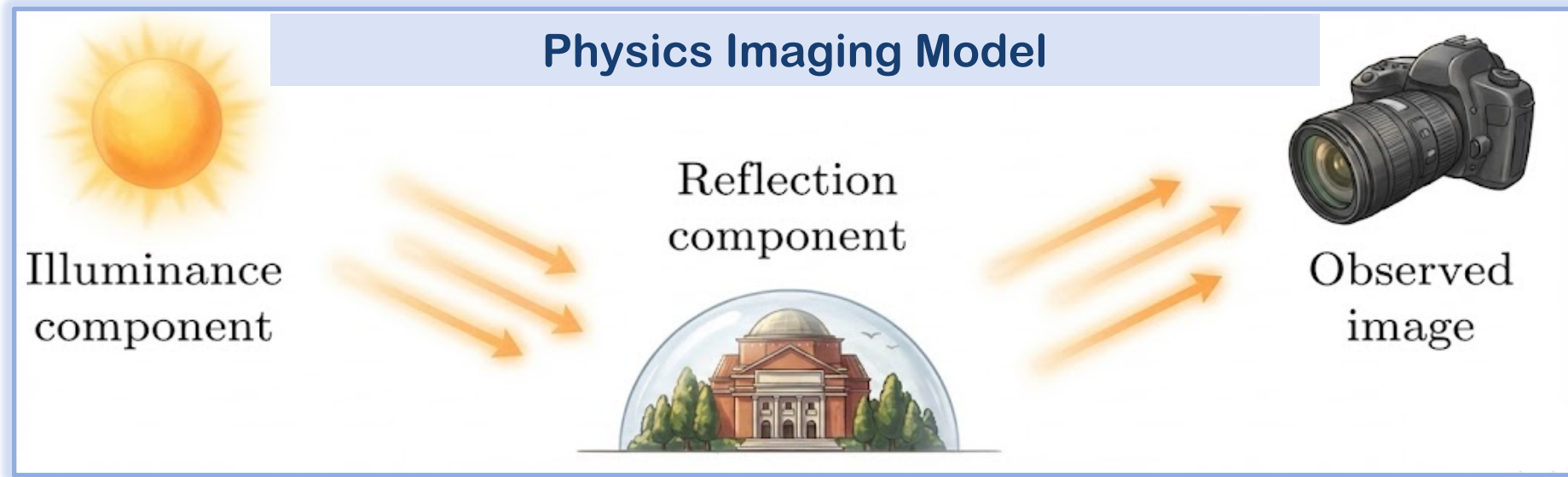


Existing methods are tailored for **specific** or **single** types of degradation, making them **ineffective** in handling more **complex composite degradations**



Existing methods **lack degradation level modeling**, causing a sharp decline in performance as degradation **intensifies**

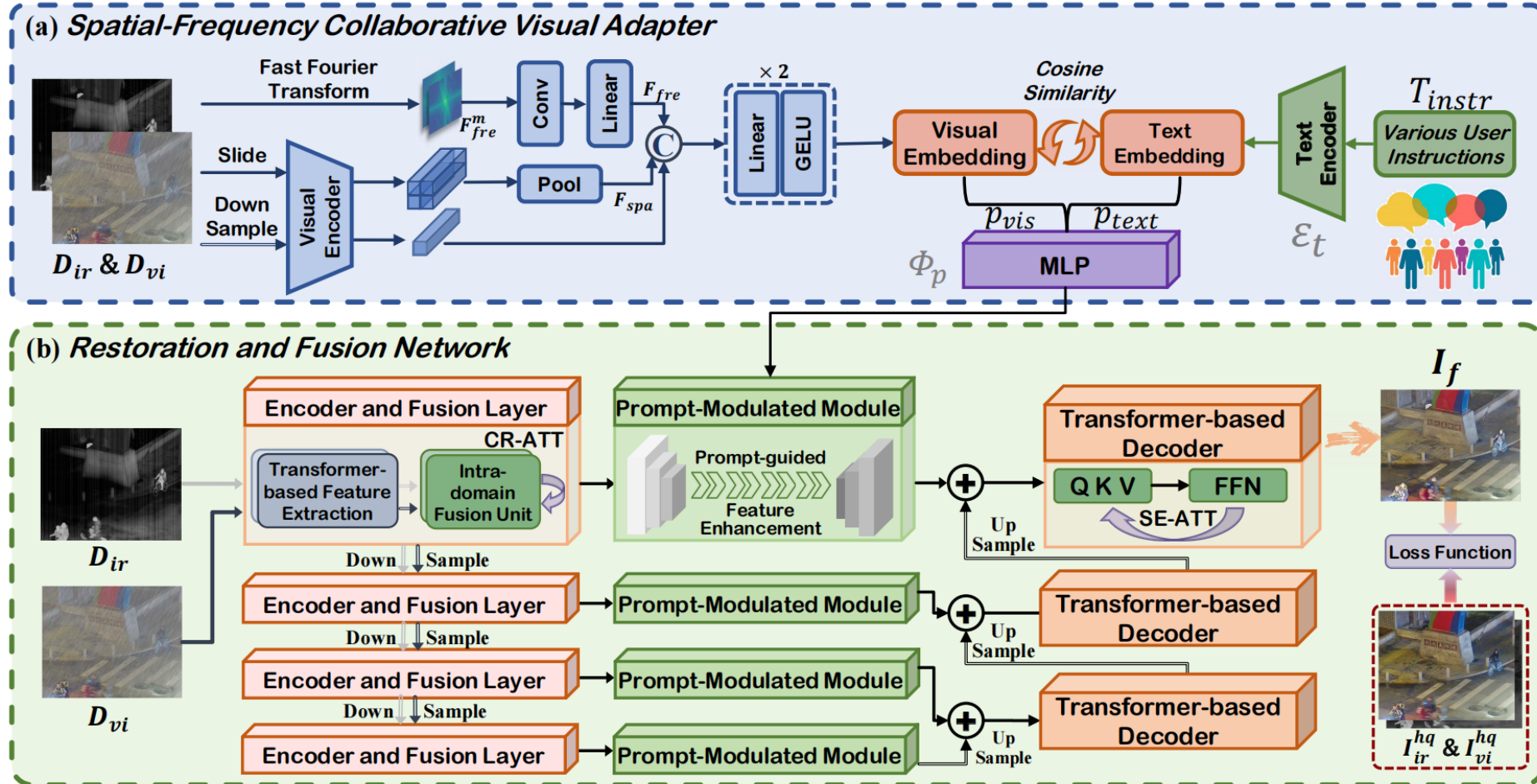
## □ Methodology — Physics-driven Degraded Imaging Model



### Physics-driven Degraded Imaging Model

- **Stripe Noise:**  $D_{ir}^s = \mathcal{P}_s(I_{ir}) = \alpha \cdot I_{ir} + \mathbf{1}_H \mathbf{n}^\top,$
- **Blur&Noise:**  $D_m^s = \mathcal{P}_s(I_m) = I_m * K(N, \theta) + \mathcal{N}(0, \sigma^2),$
- **Weather Related Degradation:**  $D_{vi}^w = \mathcal{P}_w(I_{vi}) = I_{vi} \cdot t + A(1 - t) + R,$
- **Illumination Degradation:**  $D_{vi}^i = \mathcal{P}_i(I_{vi}) = \frac{I_{vi}}{L} \cdot L^\gamma,$

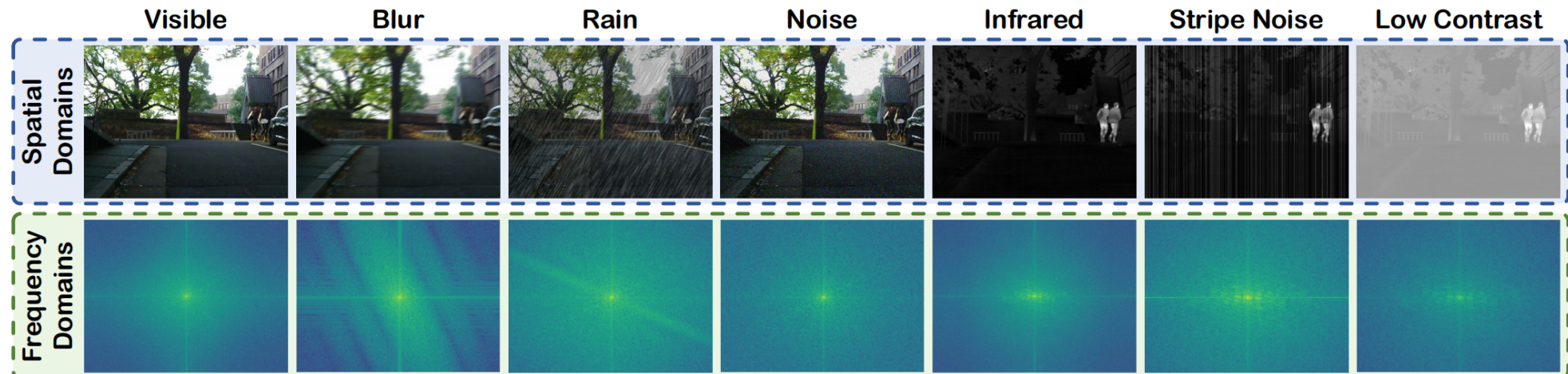
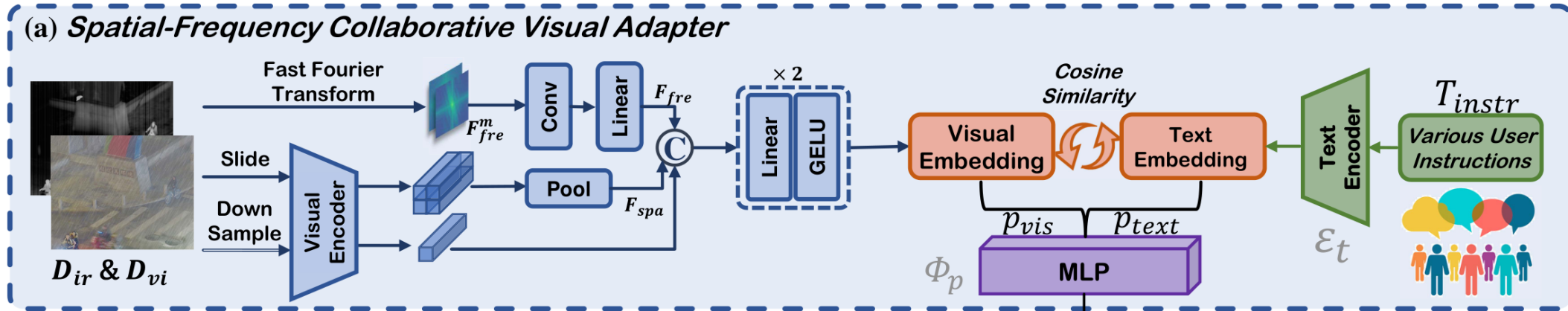
## □ Methodology — Overall Framework



Overall framework of our controllable image fusion network

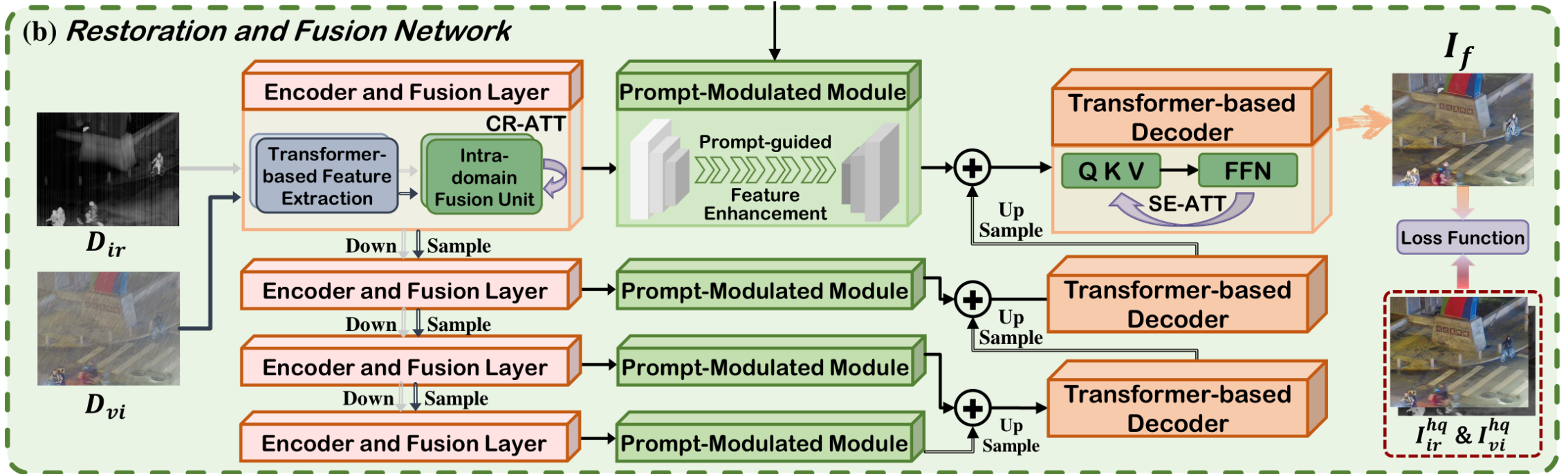


## □ Methodology — Textual-Visual Prompts Alignment



$$F_{fre}^m = \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} D_m(x, y) e^{-j2\pi(\frac{ux}{W} + \frac{vy}{H})}, F_{fre} = \text{Linear} \left( \text{Conv} \left( [F_{fre}^{ir}, F_{fre}^{vi}] \right) \right)$$

## □ Methodology — Prompt-modulated Restoration and Fusion



### Feature Aggregation

$$\{Q_{ir}, K_{ir}, V_{ir}\} = \mathcal{F}_{ir}^{qkv}(F_{ir}), \{Q_{vi}, K_{vi}, V_{vi}\} = \mathcal{F}_{vi}^{qkv}(F_{vi}).$$

$$F_f^{ir} = \text{softmax}\left(\frac{Q_{vi}K_{ir}}{\sqrt{d_k}}\right)V_{ir}, F_f^{vi} = \text{softmax}\left(\frac{Q_{ir}K_{vi}}{\sqrt{d_k}}\right)V_{vi},$$

### Feature Modulation

$$[\gamma_p, \beta_p] = \Phi_p(p)$$

$$\hat{F}_f = (1 + \gamma_p) \odot F_f + \beta_p$$

## □ Methodology — Loss Function

### Loss Function of Stage I

$$\mathcal{L}_I = \lambda_1 \underbrace{\|p_{vis} - p_{text}\|^2}_{\mathcal{L}_{mse}} + \lambda_2 \underbrace{\left(1 - \frac{p_{vis} \cdot p_{text}}{\|p_{vis}\| \|p_{text}\|}\right)}_{\mathcal{L}_{cos}},$$

✓ **MSE Loss**      ✓ **Cosine Similarity Loss**

### Loss Function of Stage II

$$\mathcal{L}_{II} = \alpha_{int} \cdot \mathcal{L}_{int} + \alpha_{ssim} \cdot \mathcal{L}_{ssim} + \alpha_{grad} \cdot \mathcal{L}_{grad} + \alpha_{color} \cdot \mathcal{L}_{color}$$

#### ✓ **Intensity Loss**

$$\mathcal{L}_{int} = \frac{1}{HW} \|I_f - \max(I_{ir}^{hq}, I_{vi}^{hq})\|_1$$

#### ✓ **SSIM Loss**

$$\mathcal{L}_{ssim} = 2 - (\text{SSIM}(I_f, I_{ir}^{hq}) + \text{SSIM}(I_f, I_{vi}^{hq}))$$

#### ✓ **Maximum Gradient Loss**

$$\mathcal{L}_{grad} = \frac{1}{HW} \|\nabla I_f - \max(\nabla I_{ir}^{hq}, \nabla I_{vi}^{hq})\|_1$$

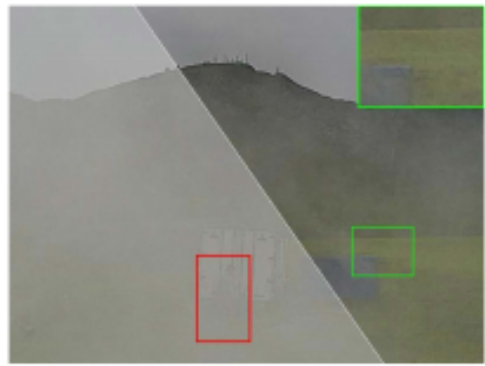
#### ✓ **Color Loss**

$$\mathcal{L}_{color} = \frac{1}{HW} \|\mathcal{F}_{CbCr}(I_f) - \mathcal{F}_{CbCr}(I_{vi}^{hq})\|_1$$



## □ Experimental Analysis — Visualized Fusion Results on Challenging Scenarios

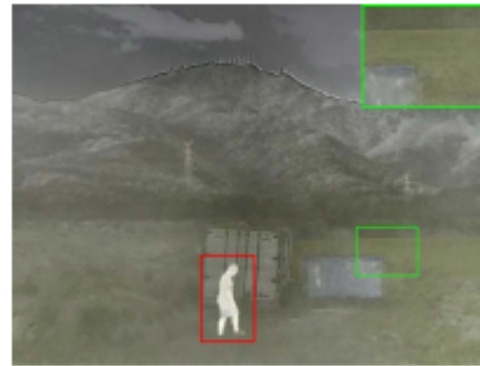
### Fusion results on real world challenging scenarios



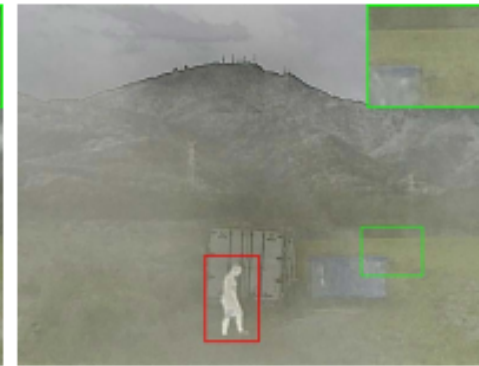
Visible  
(Deg.\ En.)



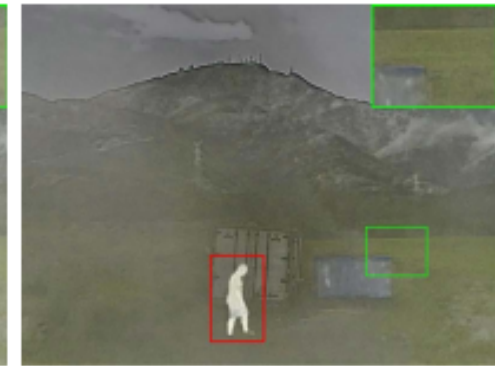
Infrared  
(Deg.\ En.)



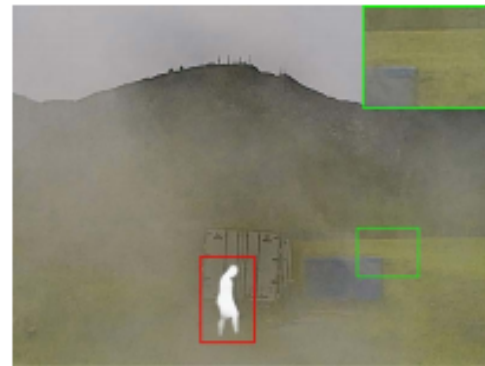
DDFM  
(ICCV 23')



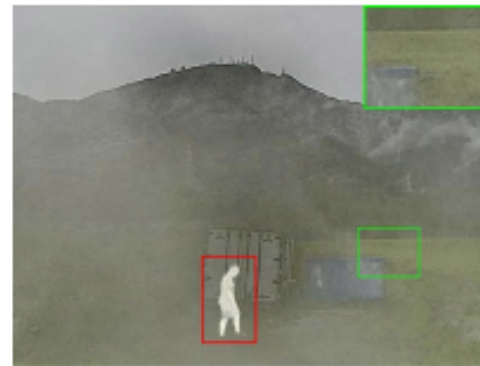
LRRNet  
(TPAMI 23')



SegMiF  
(ICCV 23')



DRMF  
(ACMMM 24')



EMMA  
(CVPR 24')



Text-IF  
(CVPR 24')



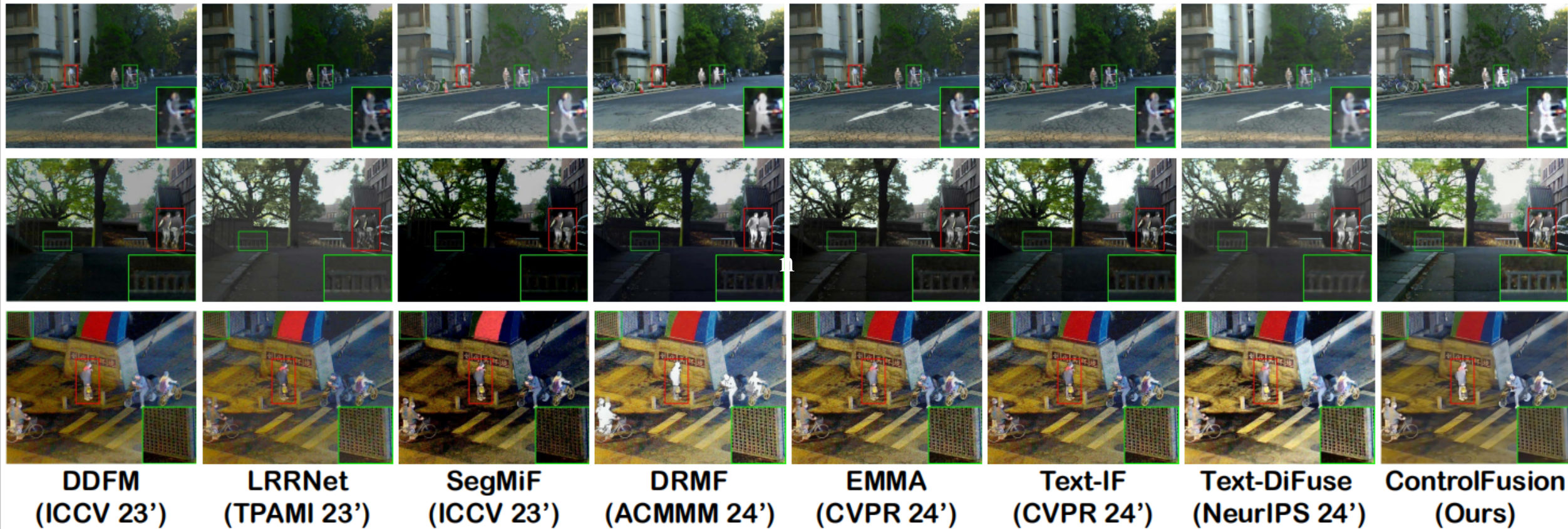
Text-DiFuse  
(Neurips 24')



ControlFusion  
(Ours)

# □ Experimental Analysis — Visualized Fusion Results on Challenging Scenarios

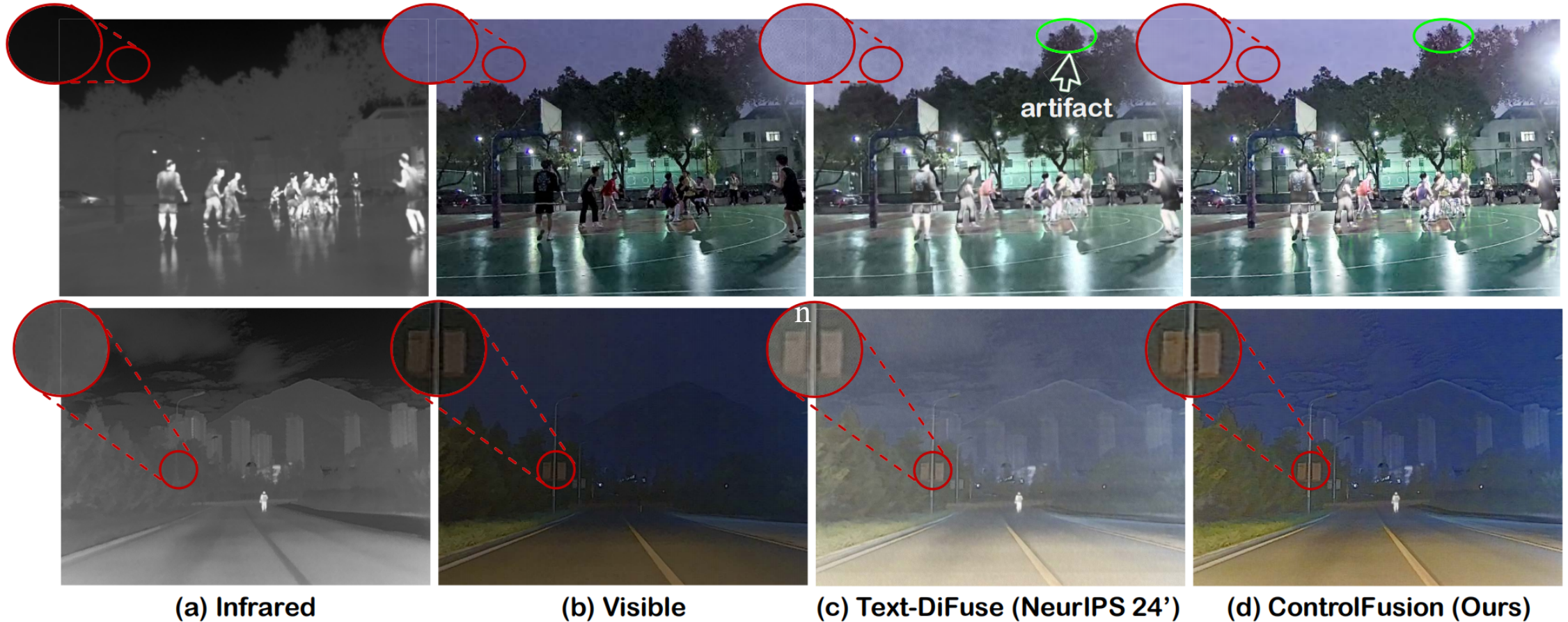
## Fusion results on real world challenging scenarios





## □ Experimental Analysis — Generalization Results on Actual Captured Images

### Generalization results on the actual captured images



## □ Experimental Analysis — Quantitative Results on Single Degradation

### Four Challenging Degradation Scenarios

- 🌀 VI (Blur): Motion or focus blur.
- 💧 VI (Rain): Weather interference.
- 🌙 VI (Low Light): Poor illumination conditions.
- ☀️ VI (Over-exposure): High dynamic range issues.

**Conclusion:** Our ControlFusion achieves State-of-the-Art performance on almost all metrics, proving superior robustness.


Methods	VI (Blur)				VI (Rain)			
	CLIP-IQA	MUSIQ	TReS	SD	CLIP-IQA	MUSIQ	TReS	SD
DDFM	0.141	39.421	39.047	35.411	0.191	38.886	46.285	36.376
DRMF	0.128	40.739	40.968	40.722	0.174	48.164	48.565	41.174
EMMA	0.131	43.472	41.744	42.553	0.185	43.924	44.916	43.378
LRRNet	0.163	43.081	37.268	45.399	0.185	43.291	41.891	46.285
SegMiF	0.152	43.005	43.516	44.000	0.195	40.528	49.094	44.274
Text-IF	0.164	44.801	46.542	48.401	0.164	41.287	47.380	49.298
Text-DiFuse	0.172	44.958	47.699	46.376	0.173	39.243	50.017	47.297
ControlFusion 🏆	0.184	47.848	50.240	50.287	0.196	52.311	52.465	50.901


Methods	VI (Low light, LL)				VI (Over-exposure, OE)			
	CLIP-IQA	MUSIQ	TReS	SD	CLIP-IQA	MUSIQ	TReS	SD
DDFM	0.156	39.495	41.782	31.759	0.143	43.167	43.440	32.099
DRMF	0.143	41.428	37.947	38.287	0.190	43.334	42.582	44.256
EMMA	0.158	39.674	44.827	40.857	0.180	46.731	47.616	40.242
LRRNet	0.164	40.486	34.836	41.639	0.160	42.548	48.414	42.190
SegMiF	0.177	41.073	46.376	44.829	0.166	49.132	38.019	38.484
Text-IF	0.163	41.096	49.174	47.287	0.172	40.298	45.999	47.330
Text-DiFuse	0.192	44.734	50.126	49.883	0.183	39.095	49.596	50.279
ControlFusion 🏆	0.183	48.420	51.072	53.787	0.191	50.301	52.961	54.218





## □ Experimental Analysis — Quantitative Results on Compound Degradations

### Compound Degradation Scenarios

 **OE + LC:** Visible Over-exposure mixed with IR Low-contrast issues.

 **LN:** Extreme Low-light conditions corrupted by Heavy Noise.

 **RH + RN:** Rain/Haze and Noise interference affecting both sensors.

 **LL + SN:** Visible Low-light coupled with IR Stripe Noise.

**Observation:** Even with dual-sensor impairments, our **ControlFusion** maintains superior stability and quality.

Methods	VI (OE) and IR (LC)				VI (Low light and Noise, LN)			
	CLIP-IQA	MUSIQ	TReS	SD	CLIP-IQA	MUSIQ	TReS	EN
DDFM	0.168	43.814	41.894	36.095	0.172	48.293	31.791	6.298
DRMF	0.184	42.399	39.374	40.847	0.201	44.363	43.063	5.875
EMMA	0.130	39.892	42.076	43.362	0.174	42.201	43.382	5.838
LRRNet	0.136	47.209	42.636	46.684	0.144	46.386	35.779	7.306
SegMiF	0.114	44.021	42.256	33.647	0.136	49.178	38.570	5.819
Text-IF	0.174	48.808	47.998	48.848	0.217	48.100	47.510	5.204
Text-DiFuse	0.131	49.021	50.980	47.640	0.185	50.775	48.610	6.440
ControlFusion	0.187	50.479	50.298	50.955	0.225	49.333	49.513	7.111

Methods	VI (RH) and IR (RN)				VI (LL) and IR (SN)			
	CLIP-IQA	MUSIQ	TReS	SD	CLIP-IQA	MUSIQ	TReS	EN
DDFM	0.151	33.440	32.134	37.342	0.189	36.433	42.630	5.776
DRMF	0.174	43.663	43.858	37.997	0.142	38.241	41.049	5.280
EMMA	0.165	39.146	44.458	51.205	0.130	37.367	43.888	6.318
LRRNet	0.128	47.954	36.831	49.917	0.154	38.426	35.970	7.007
SegMiF	0.147	42.354	39.156	31.717	0.151	41.287	37.079	6.767
Text-IF	0.158	45.821	47.626	46.543	0.140	41.429	46.220	5.525
Text-DiFuse	0.181	48.645	48.937	38.808	0.161	47.734	48.448	6.738
ControlFusion	0.179	50.107	51.091	55.417	0.167	50.632	48.971	7.055

## □ Experimental Analysis — Fusion Results with Various Level Prompts



### Dynamic Adaptation

Handles wide spectrum  
(low-light to over-exposure)



### Prompt-Driven Control

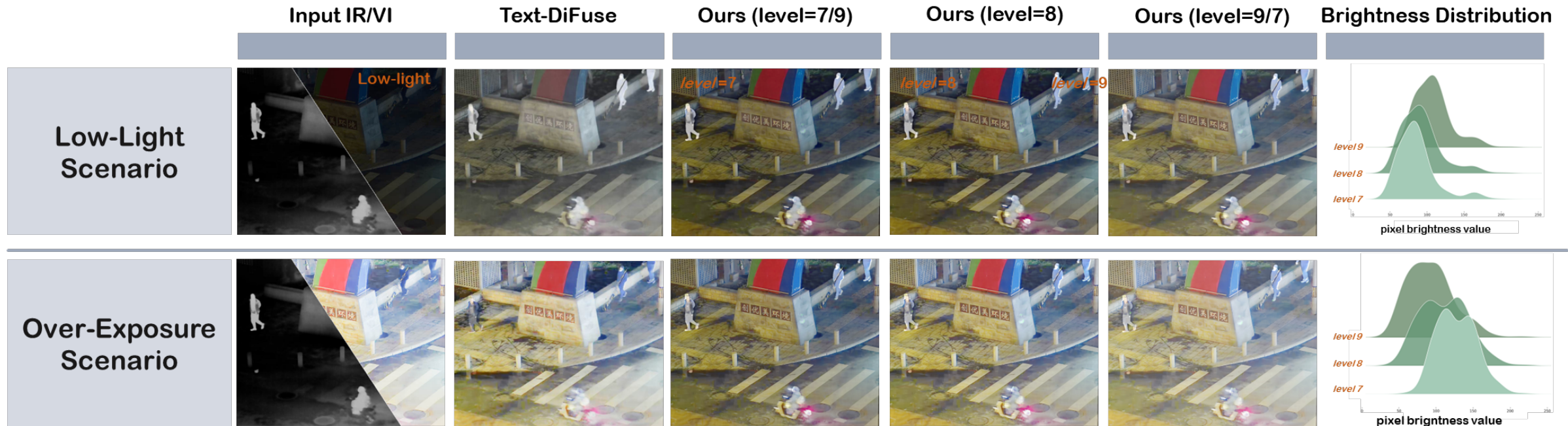
Adjust strength via specific  
prompt levels (level=7/8/9)



### Superior Robustness

high-quality fusion in  
compromised environments

**Prompt Template:** “ ... merge infrared and visible images, handling grade #level low-light/over exposure problem ... ”



## □ Experimental Analysis — Scaling-up

### ⚙ Prompt Density Scaling

Validating extensibility by increasing degradation prompt density from 2 levels to 4 levels during training.

### 📈 Performance Trend

As the diversity of prompt levels increases, the model consistently achieves higher scores across all evaluation metrics.

**Observation:** Capitalizing on richer supervision, the framework demonstrates favorable scaling properties, driven by higher prompt diversity.

Comparison of models trained with varying degradation levels in real-world datasets

Methods	FMB (Dataset 1)			LLVIP (Dataset 2)		
	2 (1,10)	2 (4,7)	4 (1,4,7,10)	2 (1,10)	2 (4,7)	4 (1,4,7,10)
CLIP-IQA	0.192	0.207	<b>0.208</b>	0.320	0.332	<b>0.347</b>
MUSIQ	52.84	52.77	<b>53.03</b>	53.50	55.36	<b>55.89</b>
TReS	63.12	63.33	<b>63.70</b>	61.89	64.38	<b>65.01</b>
SD	50.12	50.43	<b>51.71</b>	54.67	55.74	<b>56.66</b>

## □ Experimental Analysis — Quantitative Results on Practical Scenarios

**Evaluated on 4 standard datasets to ensure diverse coverage of real-world conditions**

 **MSRS & RoadScene:** Complex urban and road environments

 **LLVIP:** Challenging low-light surveillance scenarios

 **FMB:** Comprehensive multi-scenario benchmark

**Dominating Performance:** ControlFusion consistently dominates across all datasets, validating its reliability for practical deployment

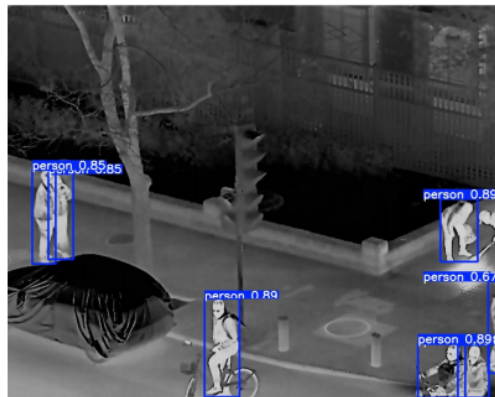
Methods	MSRS				LLVIP			
	EN	SD	VIF	Qabf	EN	SD	VIF	Qabf
DDFM	6.431	47.815	0.844	0.643	6.914	48.556	0.693	0.517
DRMF	6.268	45.117	0.669	0.550	6.901	50.736	0.786	0.626
EMMA	6.747	52.753	0.886	0.605	6.366	47.065	0.743	0.547
LRRNet	6.761	49.574	0.713	0.667	6.191	48.336	0.864	0.575
SegMiF	7.006	57.073	0.764	0.586	7.260	45.892	0.539	0.459
Text-IF	6.619	55.881	0.753	0.656	6.364	49.868	0.859	0.566
Text-DiFuse	6.990	56.698	0.850	0.603	7.546	55.725	0.883	0.659
ControlFusion	7.340	60.360	0.927	0.718	7.354	56.631	0.968	0.738

Methods	RoadScene				FMB			
	EN	SD	VIF	Qabf	EN	SD	VIF	Qabf
DDFM	6.994	47.094	0.775	0.595	6.426	40.597	0.495	0.442
DRMF	6.231	44.221	0.728	0.527	6.842	41.816	0.578	0.372
EMMA	6.959	46.749	0.698	0.664	6.788	38.174	0.542	0.436
LRRNet	7.185	46.400	0.756	0.658	6.432	48.154	0.501	0.368
SegMiF	6.736	48.975	0.629	0.584	6.363	47.398	0.539	0.482
Text-IF	6.836	47.596	0.634	0.609	7.397	47.726	0.568	0.528
Text-DiFuse	6.826	50.230	0.683	0.662	6.888	49.558	0.793	0.653
ControlFusion	7.421	51.759	0.817	0.711	7.036	50.905	0.872	0.730



## □ Experimental Analysis — Object Detection

Visualization results of target detection on the LLVIP dataset



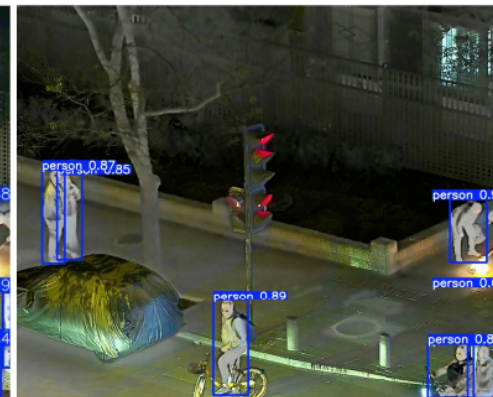
(a) Infrared



(b) DDFM



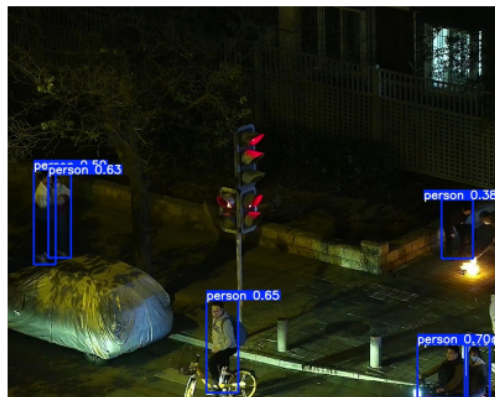
(c) DRMF



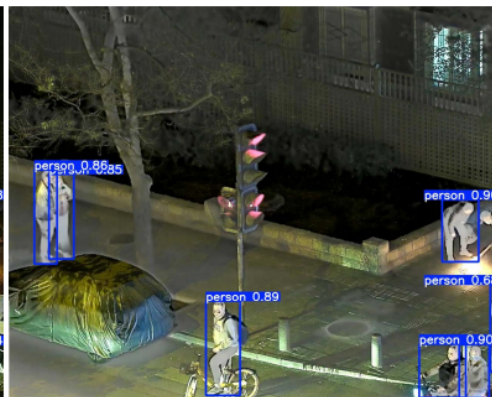
(d) EMMA



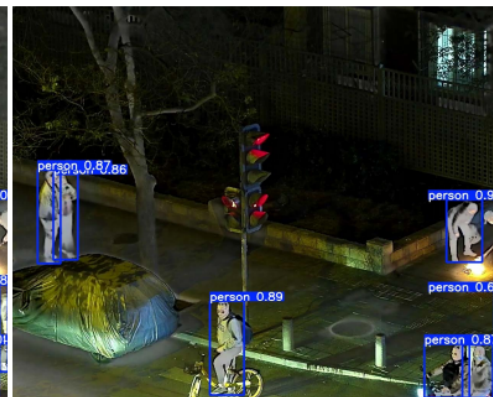
(e) LRRNet



(f) Visible



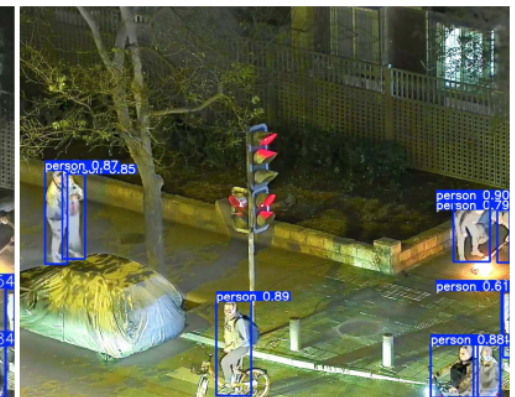
(g) SegMiF



(h) Text-IF





(i) Text-DiFuse




(j) ControlFusion

## □ Experimental Analysis — Object Detection

 **Dual-Purpose Perception:** High-quality fusion must serve both human visual preference and machine perception (downstream tasks)

 **Information Aggregation:** Effectively preserves thermal targets (from IR) and texture details (from Visible) to aid detection

 **Significant Boosting:** Achieves the highest Precision and mAP, validating the model's effectiveness in semantic understanding

**Core Value:** ControlFusion successfully bridges the gap between pixel-level fusion and high-level semantic tasks

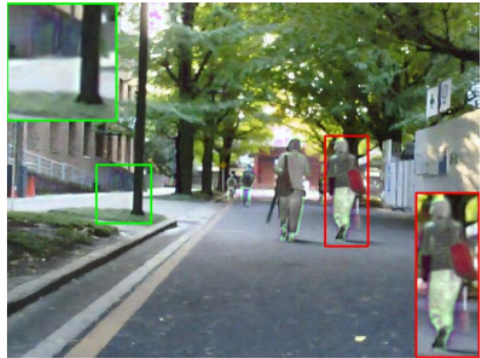
Quantitative comparison of object detection on the LLVIP dataset

Methods	Prec.	Recall	AP@0.50	AP@0.75	mAP@0.5:0.95
DDFM	0.947	0.848	0.911	0.655	0.592
DRMF	0.958	0.851	0.937	0.672	0.607
EMMA	0.942	0.872	0.927	0.647	0.598
LRRNet	0.939	0.878	0.933	0.672	0.608
SegMiF	0.965	0.896	0.931	0.690	0.603
Text-IF	0.959	0.892	0.939	0.655	0.601
Text-DiFuse	0.961	0.885	0.941	0.656	0.606
ControlFusion	0.971	0.889	0.949	0.685	0.609

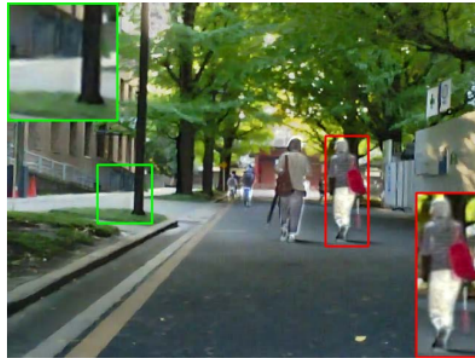


## □ Experimental Analysis — Ablation Studies

### Visual results of ablation studies under degradation scenarios



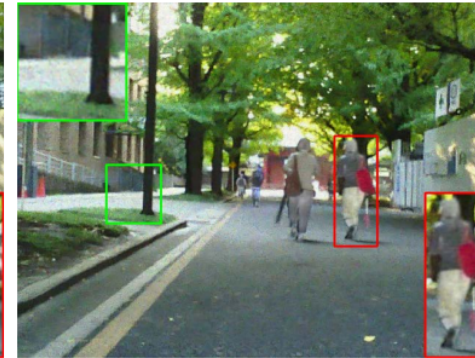
(I) w/o  $\mathcal{L}_{color}$   
(Color distortion)



(II) w/o  $\mathcal{L}_{grad}$   
(Missing textures)



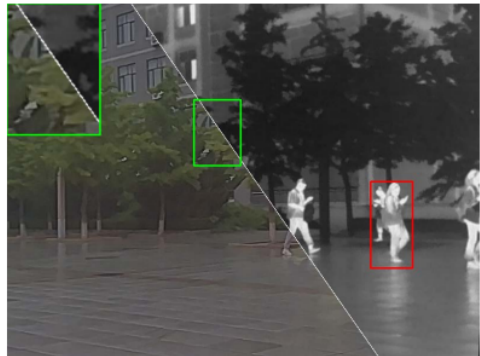
(III) w/o  $\mathcal{L}_{int}$   
(Insignificant thermal targets)



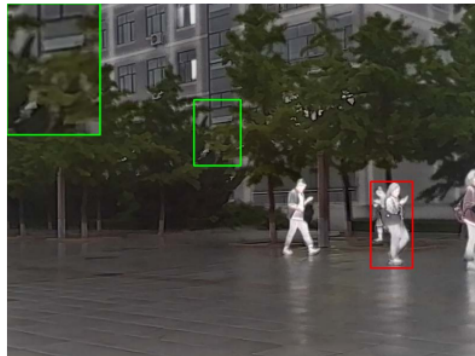
(IV) w/o PMM  
(Low quality)



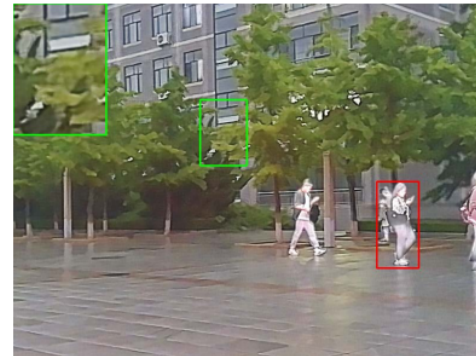
Ours  
(High quality)



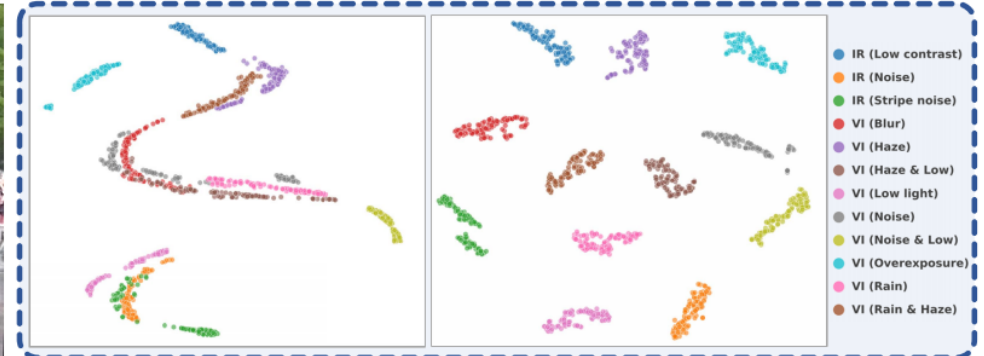
Source Images  
(Low light)



(V) w/o Frequency  
(Low quality)



Ours  
(High quality)



w/o Frequency  
(mixed)

w/ Frequency  
(Distinguishable)



## □ Experimental Analysis — Ablation Studies

📌 **Purpose:** Compare full model against five component variants to verify individual effectiveness

📋 **Ablation Variants (Configs I-V):**

- ✓ **Config I:** w/o Color Loss
- ✓ **Config II:** w/o Gradient Loss
- ✓ **Config III:** w/o Intensity Loss
- ✓ **Config IV:** w/o PMM (Prompt Modulation Module)
- ✓ **Config V:** w/o Frequency Branch

Table 6: Quantitative results of the ablation studies.

Configs	VI(LL & Noise)				VI(OE) and IR(LC)				VI (RH) and IR(Noise)			
	CLIP-IQA	MUSIQ	TReS	EN	CLIP-IQA	MUSIQ	TReS	SD	CLIP-IQA	MUSIQ	TReS	SD
I	0.132	42.424	42.839	4.788	0.166	41.983	42.841	39.917	0.147	43.619	47.932	48.935
II	0.152	45.582	45.358	5.855	0.151	43.646	42.002	39.208	0.167	47.862	45.007	44.347
III	0.154	46.571	44.495	5.013	0.155	44.286	44.561	42.068	0.156	43.007	43.816	41.544
IV	0.129	38.960	41.310	5.414	0.172	41.743	39.125	38.748	0.118	48.882	46.245	45.910
V	0.173	45.281	46.291	6.279	0.181	45.386	47.519	46.860	0.149	46.714	48.094	46.950
Ours	0.225	49.333	49.513	7.111	0.187	50.479	50.298	50.955	0.179	50.107	51.091	55.417

## □ Conclusion & Future Work



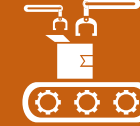
### Physics-Driven Model

- Integrates **Retinex theory** and **atmospheric scattering** principles
- Bridges the gap between **synthetic** data and **real-world** images
- Specifically models the degradation of **infrared-visible dual modalities**



### Controllable Framework

- Propose **ControlFusion**, a unified framework using prompts as a medium
- Uniformly models diverse degradation **types and degrees**
- **Controllability**: Responds precisely to user-specific customization needs.

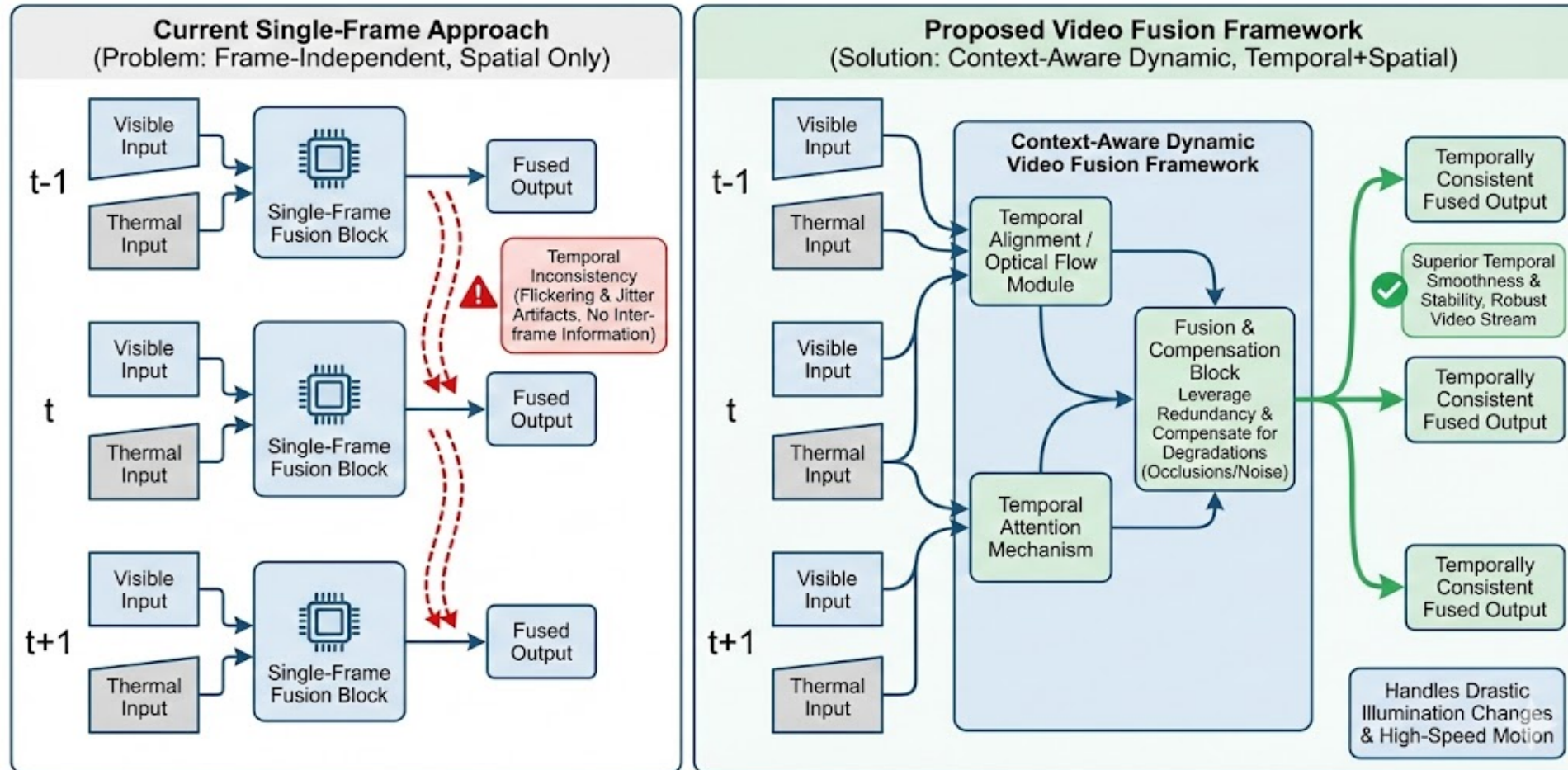


### Automated Deployment

- Devised a novel **visual adapter** to integrate frequency characteristics
- Directly extracts text-aligned **degradation prompts** from input images
- Enables **automated deployment** without manual intervention

## □ Conclusion & Future Work

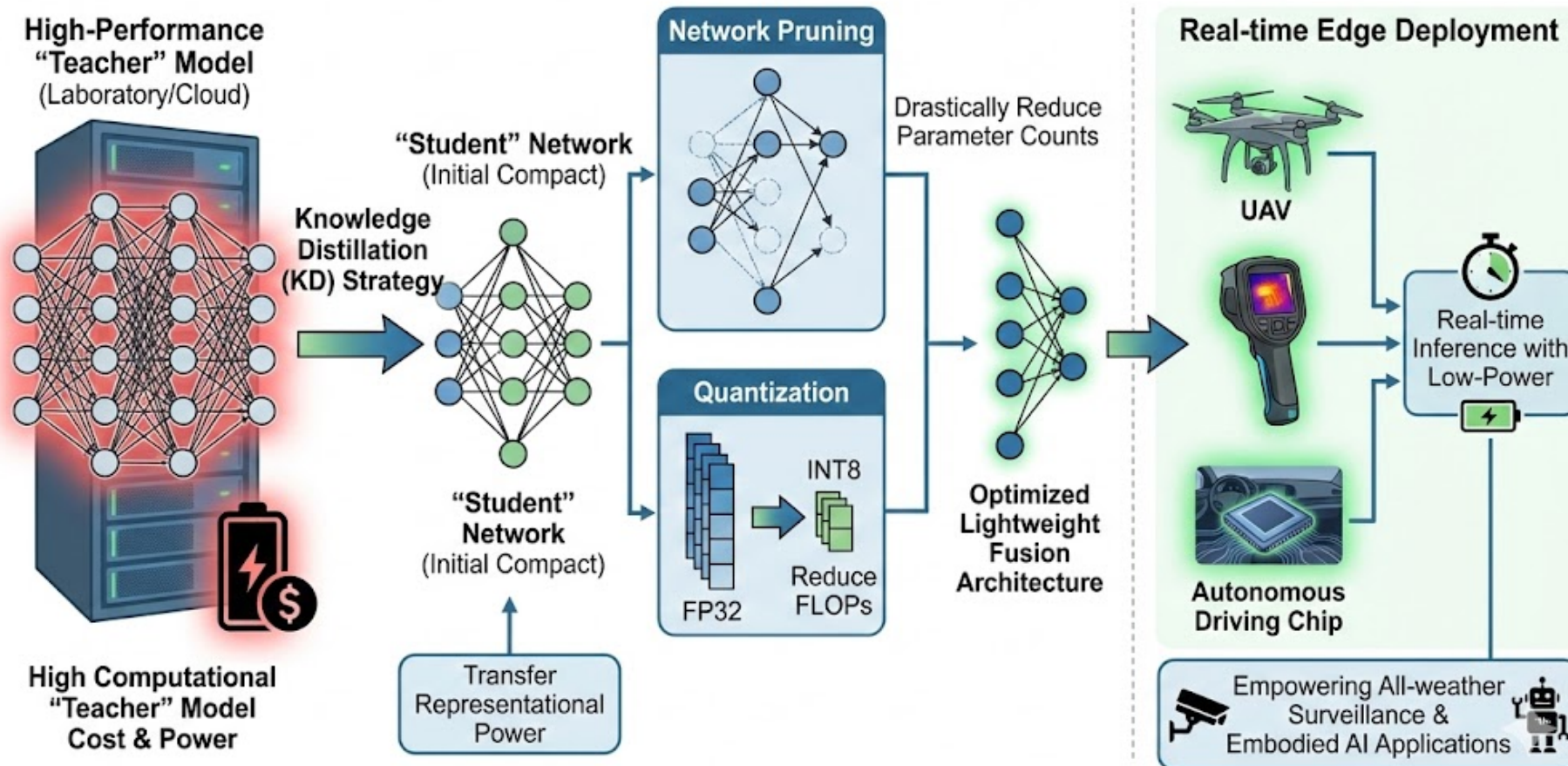
### Extension to Video Fusion with Temporal Consistency



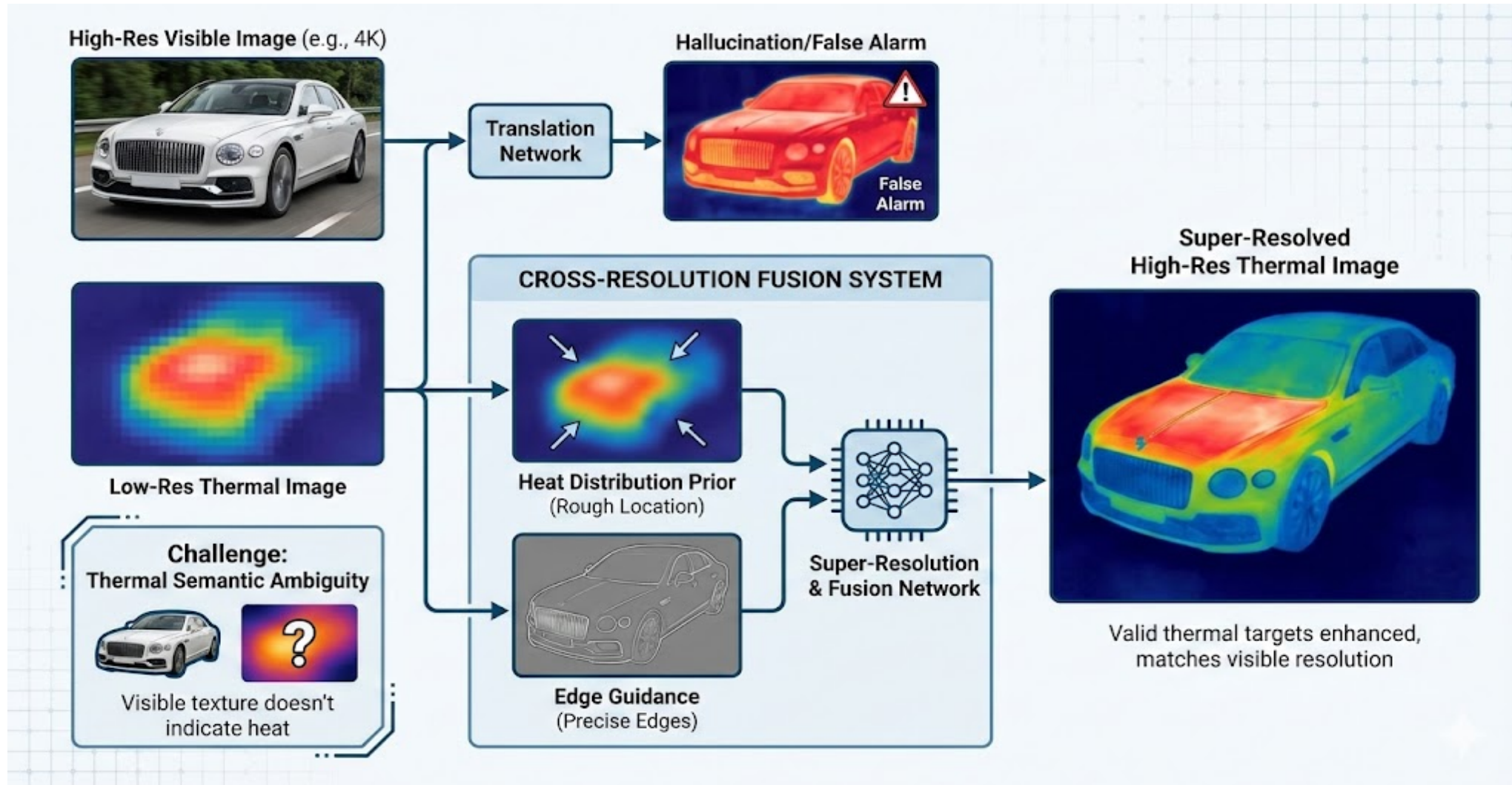


## □ Conclusion & Future Work

### Lightweight Architecture for Real-time Edge Deployment



## □ Conclusion & Future Work





# ControlFusion: A Controllable Image Fusion Network with Language-Vision Degradation Prompts

Thank You for Watching!

 View on Github: <https://github.com/Linfeng-Tang/ControlFusion>

This work was supported by National Natural Science Foundation of China (No. 62276192).

