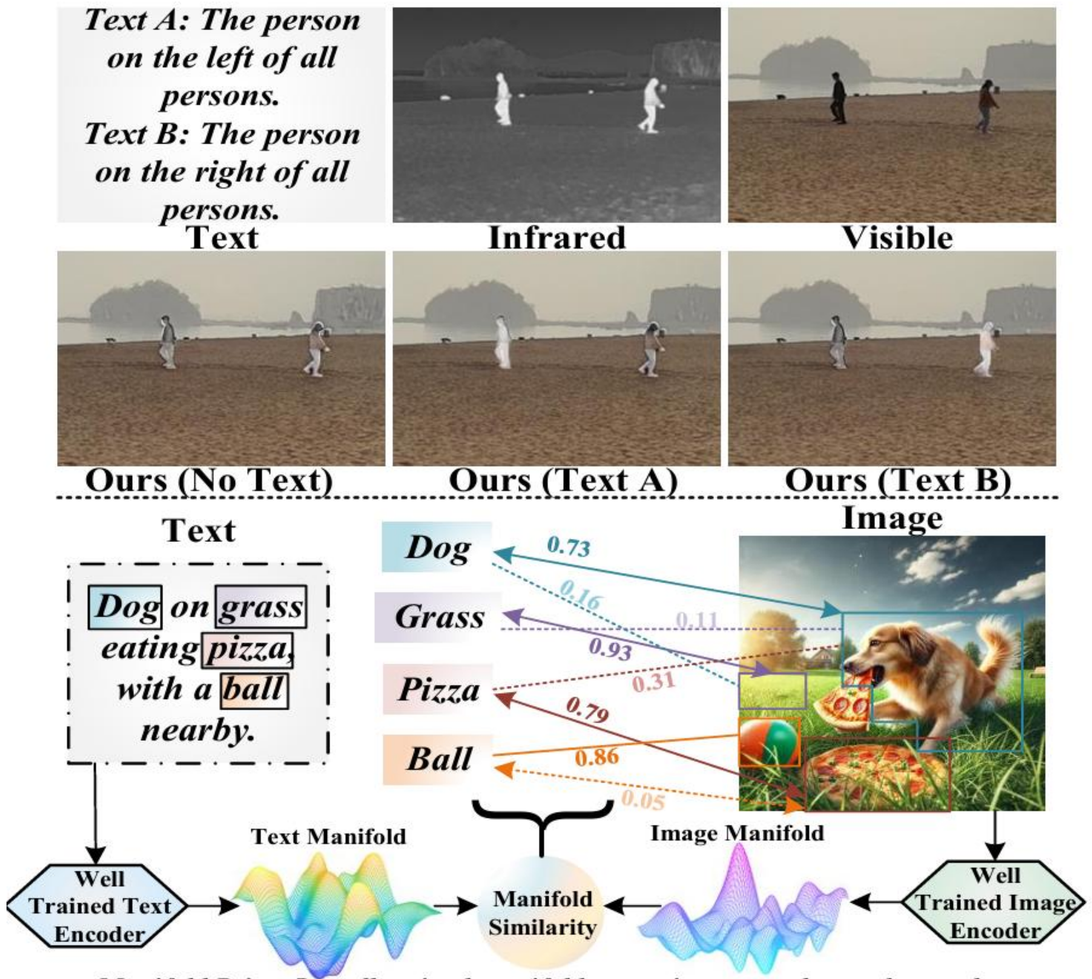
Highlight What You Want: Weakly-Supervised Instance-Level Controllable Infrared-Visible Image Fusion

Zeyu Wang[1], Jizheng Zhang[2], Haiyu Song[1*], Mingyu Ge[1], Jiayu Wang[1], Haoran Duan[3*]

[1]Dalian Minzu University [2]University of Macau [3]Tsinghua University

Background & Motivation



Manifold Prior: In well-trained manifold space, image patches and text tokens describing the same object exhibit high manifold similarity, while unrelated pairs do not.

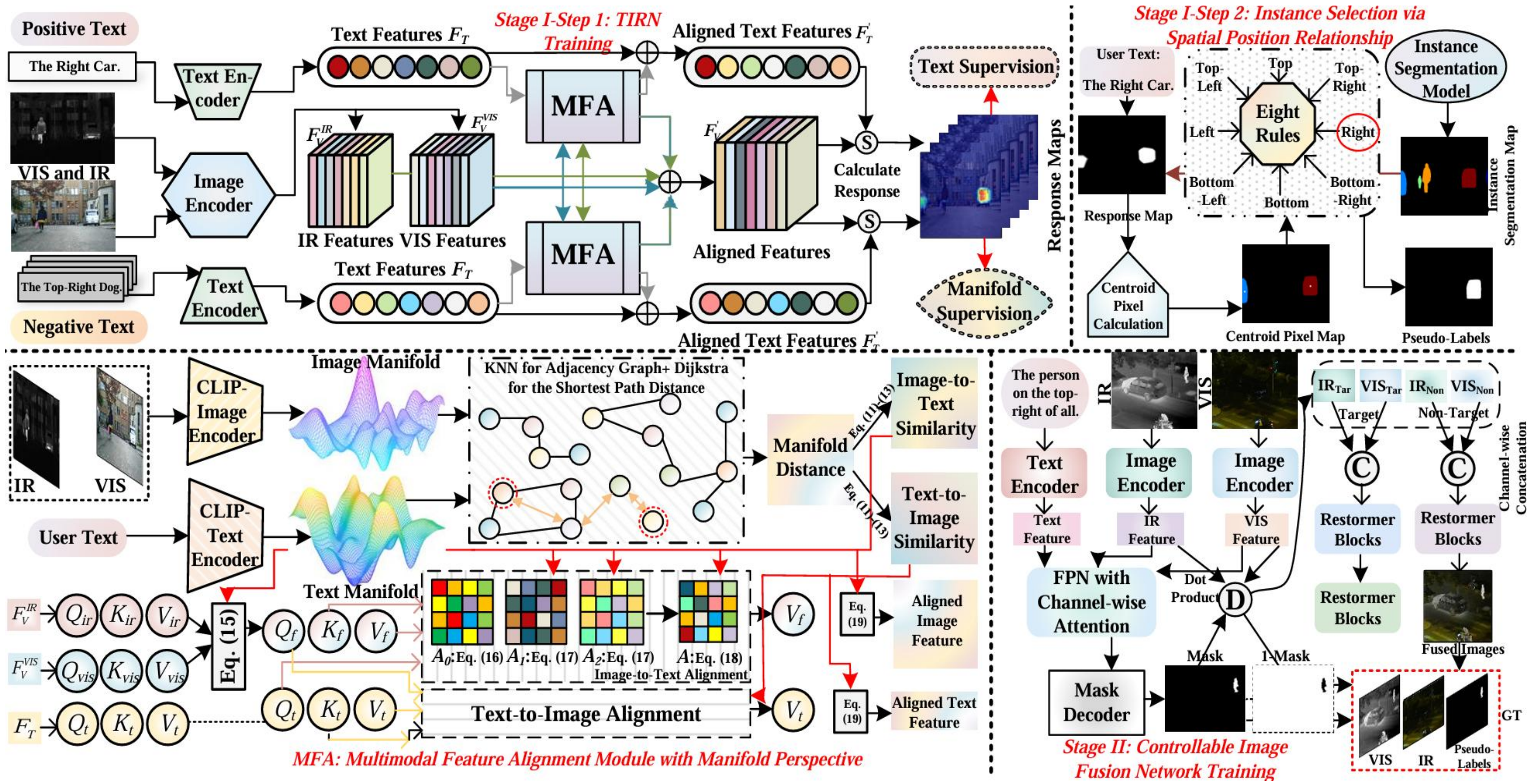
- IR: robust in low light, lacks texture;

  VIS: rich details, light-dependent.

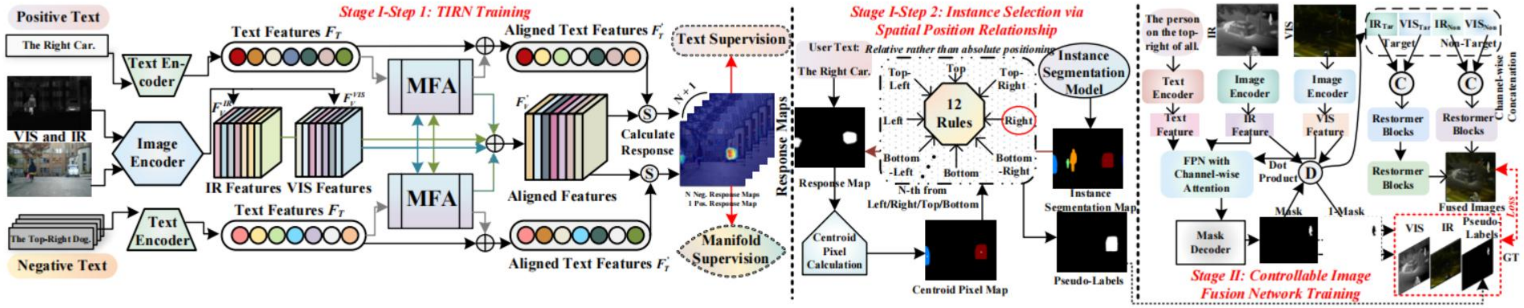
  Fusion aims to combine both.
- Limitation: Most methods are uncontrollable; controllable ones are semantic-level only.
- Need: Emphasize a user-referred instance via natural language in real scenarios

Contributions

- Instance-level controllable fusion from natural-language prompts.
- Two-stage weak supervision:
  - Stage I: Pseudo-labels via TIRN + MFA (text–image manifold similarity); ISM with 12 spatial rules.
  - Stage II: Fusion network with target vs. non-target region strategies.
- Leverage a text–image manifold prior for alignment.



Framework Overview



Stage I (Pseudo-labels):

Input IR, VIS, text → TIRN response maps.

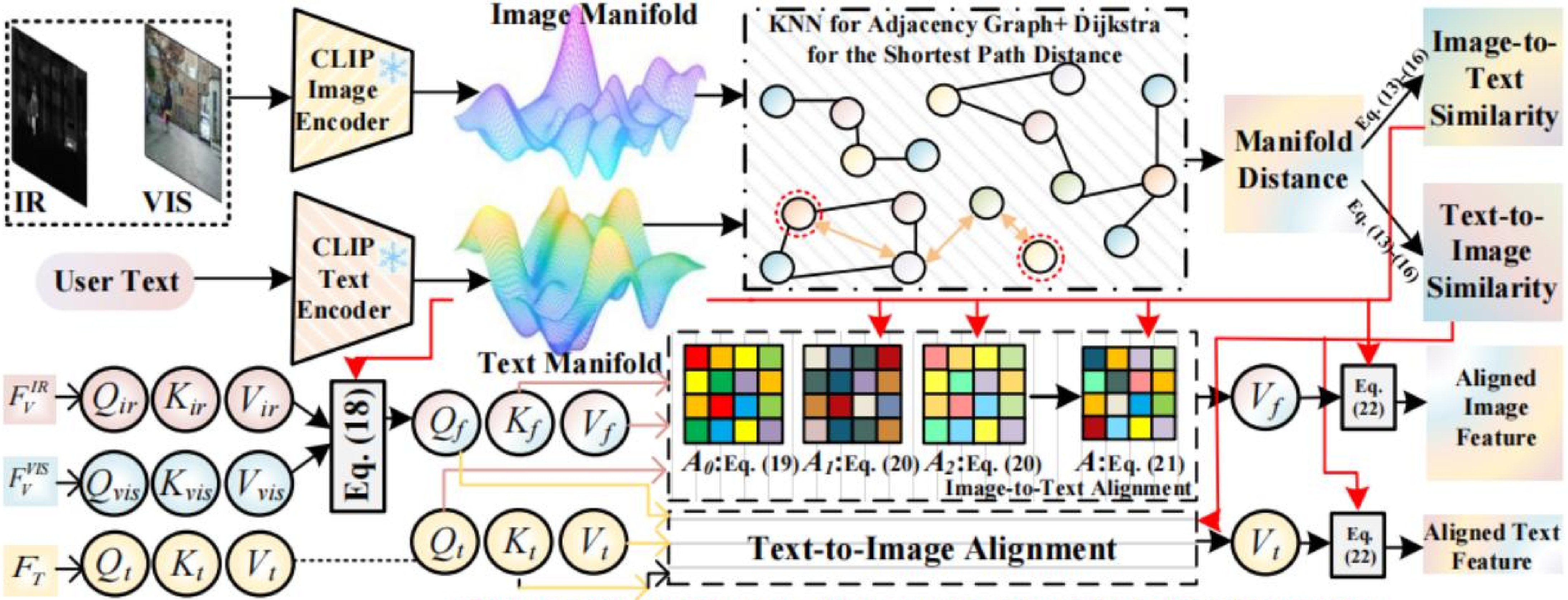
ISM picks the referred instance using relative rules (left/right/top-right/N-th...).

Stage II (Fusion):

Train with pseudo-labels; enhance target (IR luminance + VIS color/texture), preserve non-target quality.

Inference runs only the fusion network to localize & fuse conditioned on text.

Manifold-Based Feature Alignment Module



MFA: Multimodal Feature Alignment Module with Manifold Perspective

Qualitative Results

Quantitative results. Each image pair is given a text description that randomly refers to an object in the image.

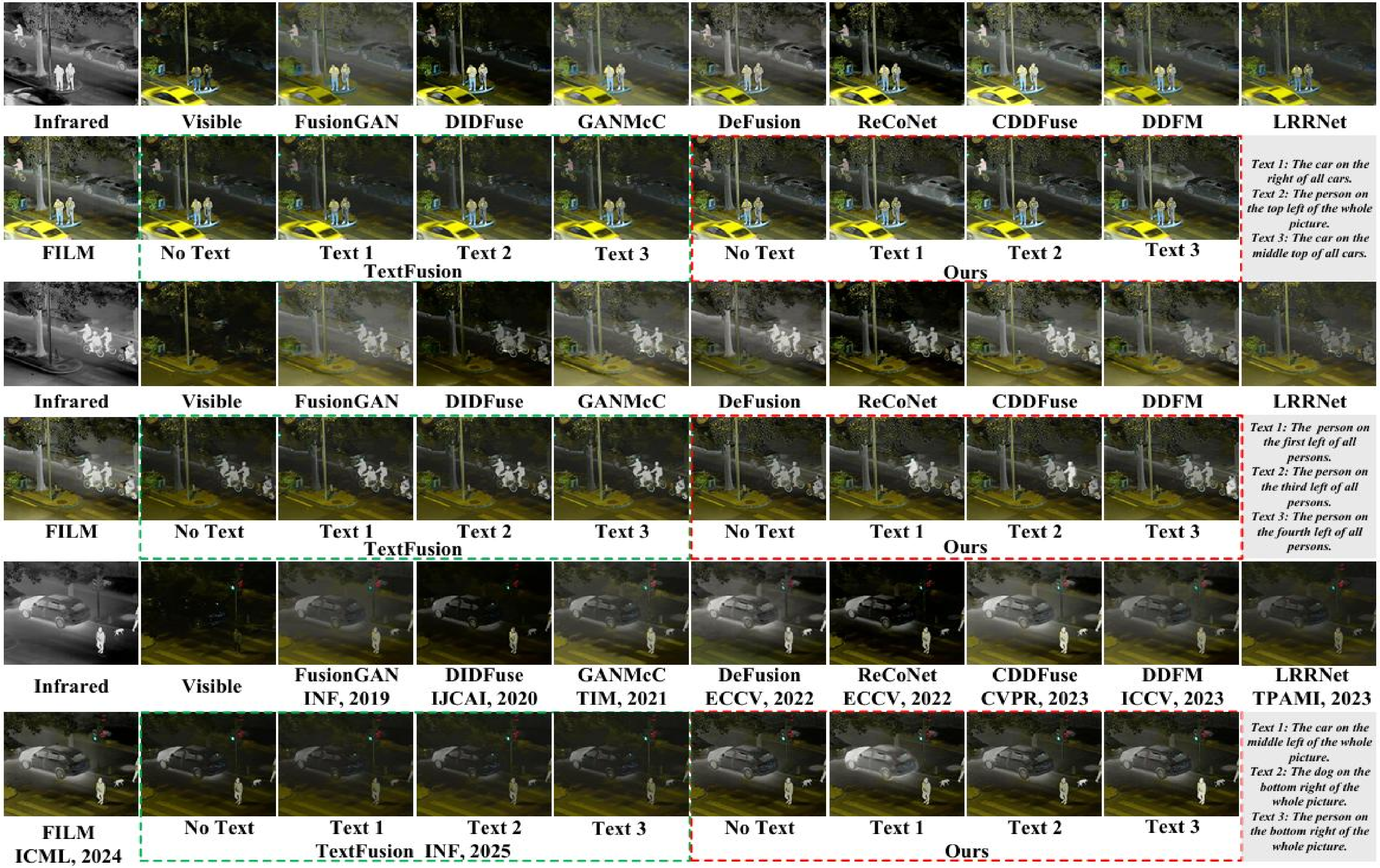
Test Sets	Methods	$Q_{CB}\uparrow$	$Q_Y\uparrow$	$Q_E\uparrow$	$Q_W\uparrow$	$SF\uparrow$	$AG\uparrow$	$VIF\uparrow$	$Q_{AB/F}\uparrow$
LLVIP	FusionGAN	0.251	0.487	0.306	0.289	5.970	1.609	0.498	0.212
	DIDFuse	0.424	0.338	0.432	0.389	10.313	2.098	0.444	0.267
	GANMcC	0.326	0.572	0.368	0.413	6.286	1.874	0.654	0.292
	DeFusion	0.432	0.734	0.576	0.650	8.940	2.523	0.776	0.466
	ReCoNet	0.407	0.395	0.611	0.551	9.915	2.774	0.580	0.408
	CDDFuse	0.459	0.857	0.767	0.762	13.278	3.432	0.958	0.641
	DDFM	0.406	0.613	0.370	0.451	5.785	1.869	0.714	0.300
	LRRNet	0.360	0.521	0.512	0.411	8.874	2.277	0.558	0.405
	FILM	0.485	0.759	0.825	0.791	14.361	3.926	0.976	0.675
	TextFusion	0.455	0.692	0.654	0.552	11.765	2.894	0.738	0.495
	Ours	0.511	0.784	0.843	0.803	14.572	4.005	0.996	0.711
MSRS	FusionGAN	0.322	0.391	0.206	0.206	4.354	1.446	0.442	0.140
	DIDFuse	0.407	0.238	0.432	0.402	9.644	2.006	0.304	0.202
	GANMcC	0.424	0.574	0.404	0.444	5.664	1.999	0.635	0.302
	DeFusion	0.514	0.749	0.730	0.753	8.146	2.644	0.730	0.507
	ReCoNet	0.378	0.347	0.720	0.692	9.975	2.990	0.490	0.404
	CDDFuse	0.567	0.827	0.868	0.859	11.556	3.734	1.051	0.693
	DDFM	0.482	0.662	0.582	0.579	7.388	2.513	0.743	0.474
	LRRNet	0.393	0.507	0.659	0.632	8.473	2.641	0.541	0.454
	FILM	0.569	0.822	0.873	0.863	11.726	3.858	1.056	0.723
	TextFusion	0.490	0.618	0.795	0.805	10.230	3.045	0.694	0.475
	Ours	0.570	0.875	0.887	0.876	11.788	3.877	0.997	0.698

Test Sets	Methods	$Q_{CB}\uparrow$	$Q_Y\uparrow$	$Q_E\uparrow$	$Q_W\uparrow$	$SF\uparrow$	$AG\uparrow$	$VIF\uparrow$	$Q_{AB/F}\uparrow$
M^3FD	FusionGAN	0.330	0.552	0.376	0.333	8.012	2.688	0.388	0.264
	DIDFuse	0.445	0.653	0.668	0.654	15.423	5.306	0.669	0.511
	GANMcC	0.392	0.615	0.409	0.431	7.590	2.695	0.527	0.289
	DeFusion	0.427	0.672	0.502	0.467	8.408	2.945	0.549	0.345
	ReCoNet	0.455	0.740	0.706	0.670	11.955	4.479	0.604	0.508
	CDDFuse	0.475	0.870	0.793	0.722	16.491	5.417	0.781	0.632
	DDFM	0.405	0.656	0.555	0.532	9.725	3.379	0.606	0.449
	LRRNet	0.435	0.721	0.646	0.574	26.690	9.269	0.762	0.461
	FILM	0.493	0.839	0.816	0.748	16.757	5.545	0.806	0.653
	TextFusion	0.477	0.713	0.755	0.662	16.836	5.523	0.619	0.544
	Ours	0.517	0.905	0.833	0.754	16.627	5.609	0.838	0.694
TNO	FusionGAN	0.408	0.539	0.386	0.373	6.269	2.362	0.418	0.224
	DIDFuse	0.463	0.655	0.602	0.614	11.768	4.249	0.593	0.403
	GANMcC	0.437	0.607	0.420	0.445	6.217	2.513	0.513	0.275
	DeFusion	0.482	0.707	0.592	0.568	6.598	2.675	0.553	0.359
	ReCoNet	0.463	0.673	0.592	0.610	7.958	3.353	0.531	0.373
	CDDFuse	0.450	0.787	0.697	0.659	11.621	4.330	0.730	0.496
	DDFM	0.437	0.485	0.473	0.484	8.128	3.213	0.371	0.292
	LRRNet	0.483	0.701	0.513	0.513	9.438	3.627	0.538	0.367
	FILM	0.483	0.821	0.745	0.726	12.579	4.556	0.725	0.529
	TextFusion	0.525	0.750	0.581	0.582	10.217	3.986	0.598	0.392
	Ours	0.520	0.829	0.746	0.733	12.935	4.791	0.704	0.566

Qualitative comparison

ICCV HONOLULU HAWAII OCT 19-23, 2025

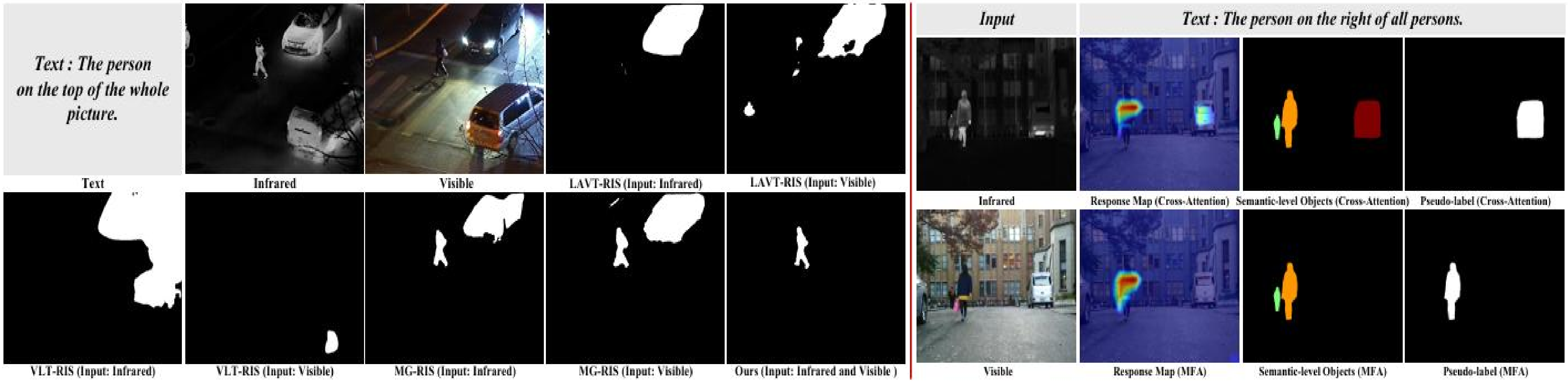
The first nine UIF models yield static results regardless of text. TextFusion and ours take one text and two source images as input. TextFusion highlights semantic-level objects, while our model targets the referenced instance.



Qualitative Results

Validation of the Necessity of Tailored Instance Localization Method for VIS-IR.

The left part shows a comparison of our localization method with recent RIS models on VIS and IR images. The right part illustrates the impact of our Manifold-Based Feature Alignment Module and cross-attention on pseudo-label accuracy



Downstream Task: Targeted Object Detection

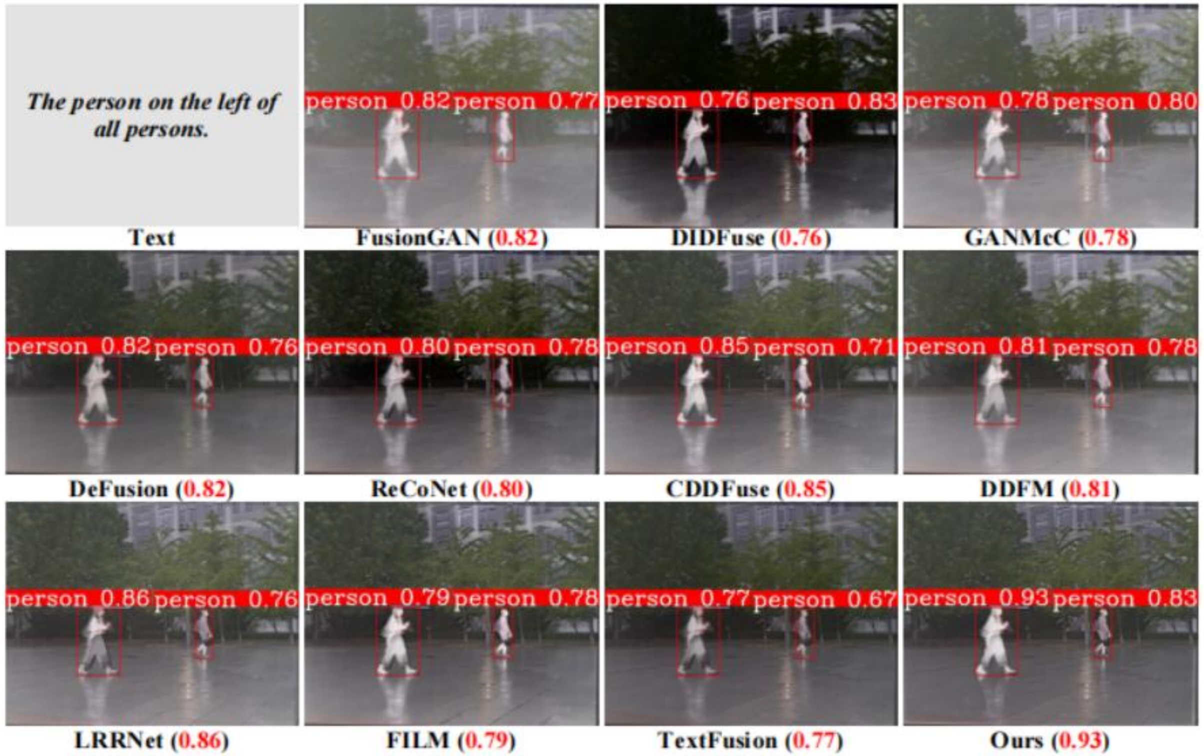


Table 3. Comparison results of fusion models on the TOD task.

Fusion Methods	Recall	mAP @0.50	mAP @0.75	mAP @[0.5:0.95]
FusionGAN [31]	0.207	0.307	0.179	0.170
DIDFuse [55]	0.272	0.402	0.250	0.231
GANMcC [32]	0.267	0.393	0.235	0.222
DeFusion [23]	0.259	0.378	0.233	0.214
ReCoNet [11]	0.287	0.423	0.252	0.243
CDDFuse [56]	0.292	0.430	0.261	0.246
DDFM [57]	0.281	0.422	0.245	0.234
LRRNet [16]	0.304	0.447	0.260	0.252
FILM [59]	0.300	0.45	0.261	0.251
TextFusion [4]	0.288	0.423	0.253	0.243
Ours	0.314	0.472	0.268	0.263

Our model enhances TOD by highlighting referenced instances.
The confidence of the person instance described by the text in Yolo detection is higher than that of other algorithms

Thanks For Listening!