Miss-ReID: Delivering Robust Multi-Modality Object Re-Identification Despite Missing Modalities

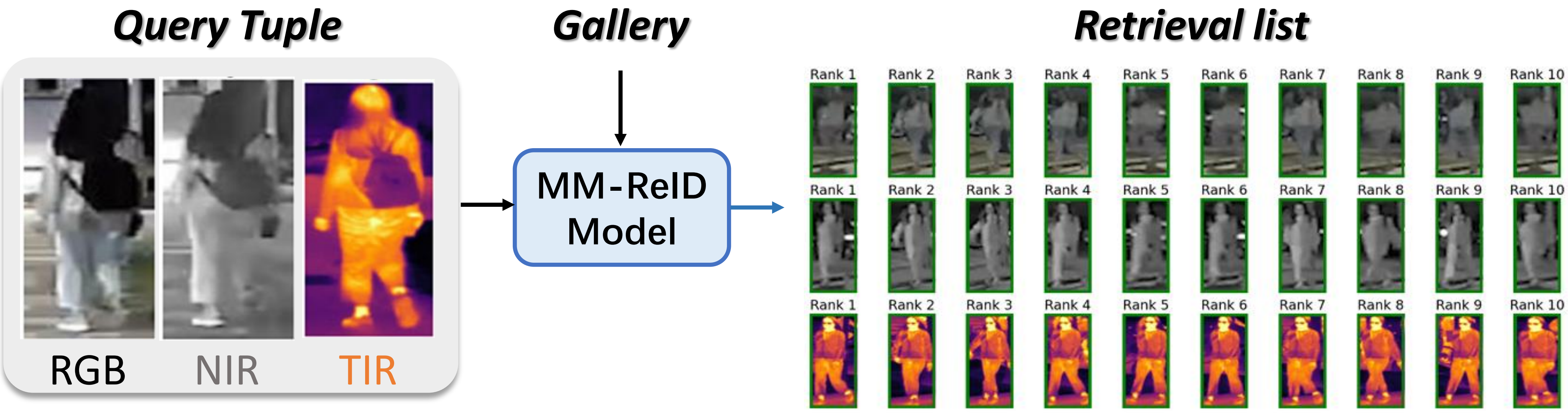
Ruida Xi

State Key Laboratory of Electromechanical Integrated
Manufacturing of High-Performance Electronic Equipment, Xidian University

Paper: https://nips.cc/virtual/2025/poster/119663

Background

Multi-Modality Object Re-Identification

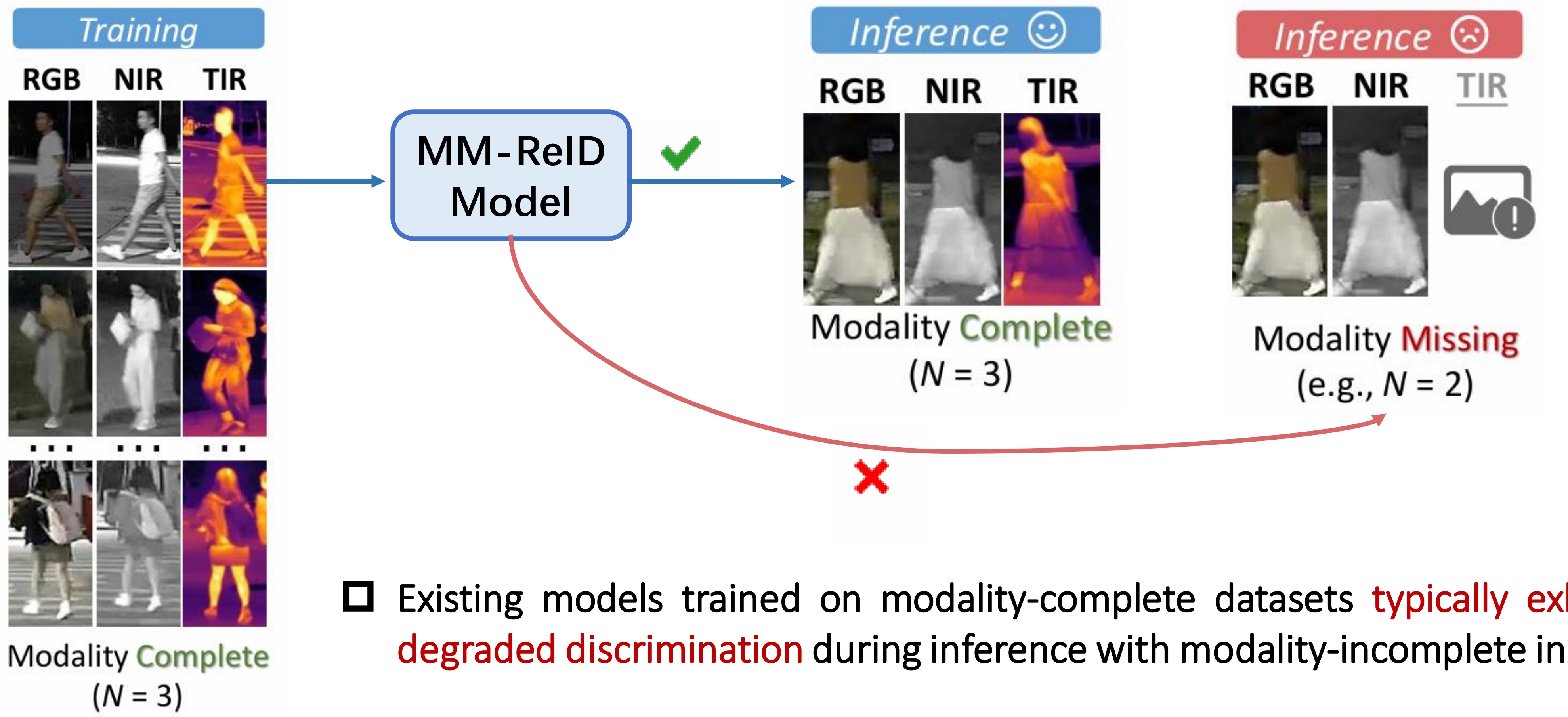


Multi-modality object Re-IDentification (ReID) targets to retrieve special objects by integrating complementary information from diverse visual sources.

[1] Wang Y, Liu Y, Zheng A, et al. Decoupled feature-based mixture of experts for multi-modal object re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(8): 8141-8149.

Problem

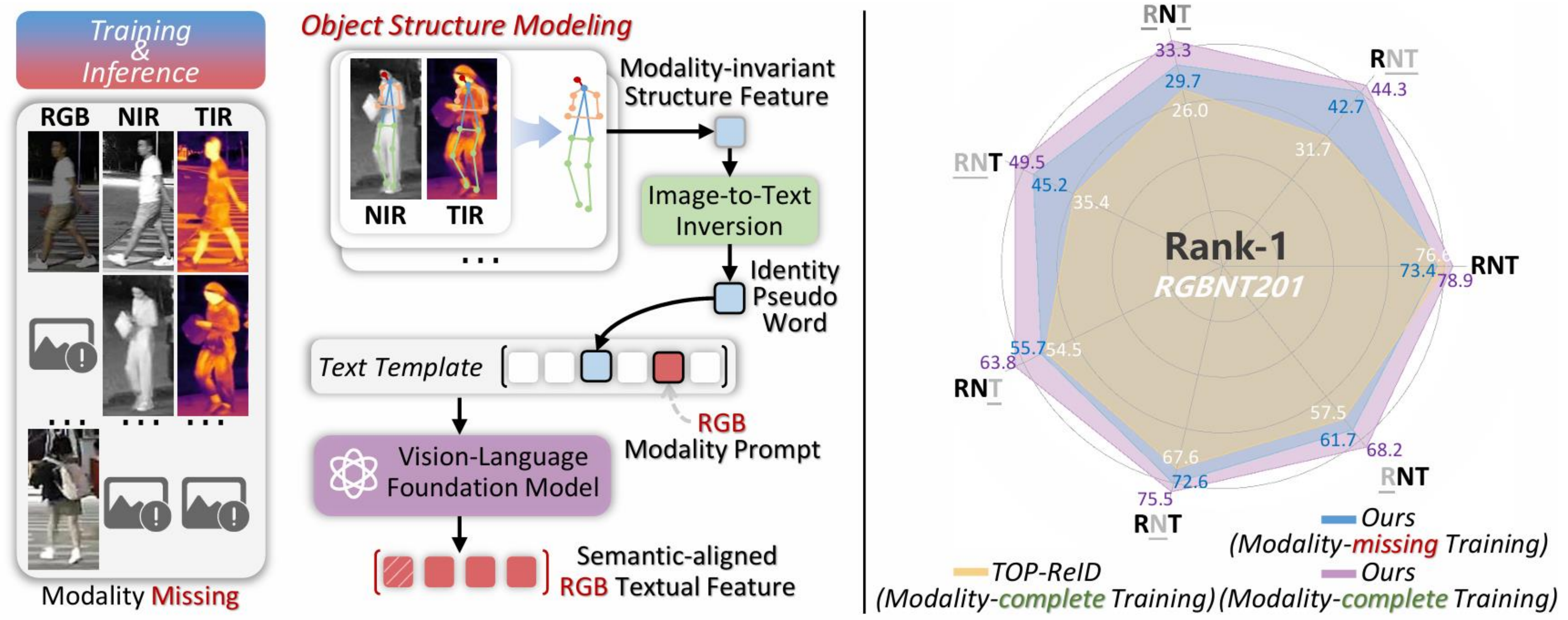
- Though exhibiting promising performance, existing multi-modality ReID methods typically rely on an assumption regarding the modality completeness, which may not hold in practice owing to privacy protections, sensor failures or security requirements.



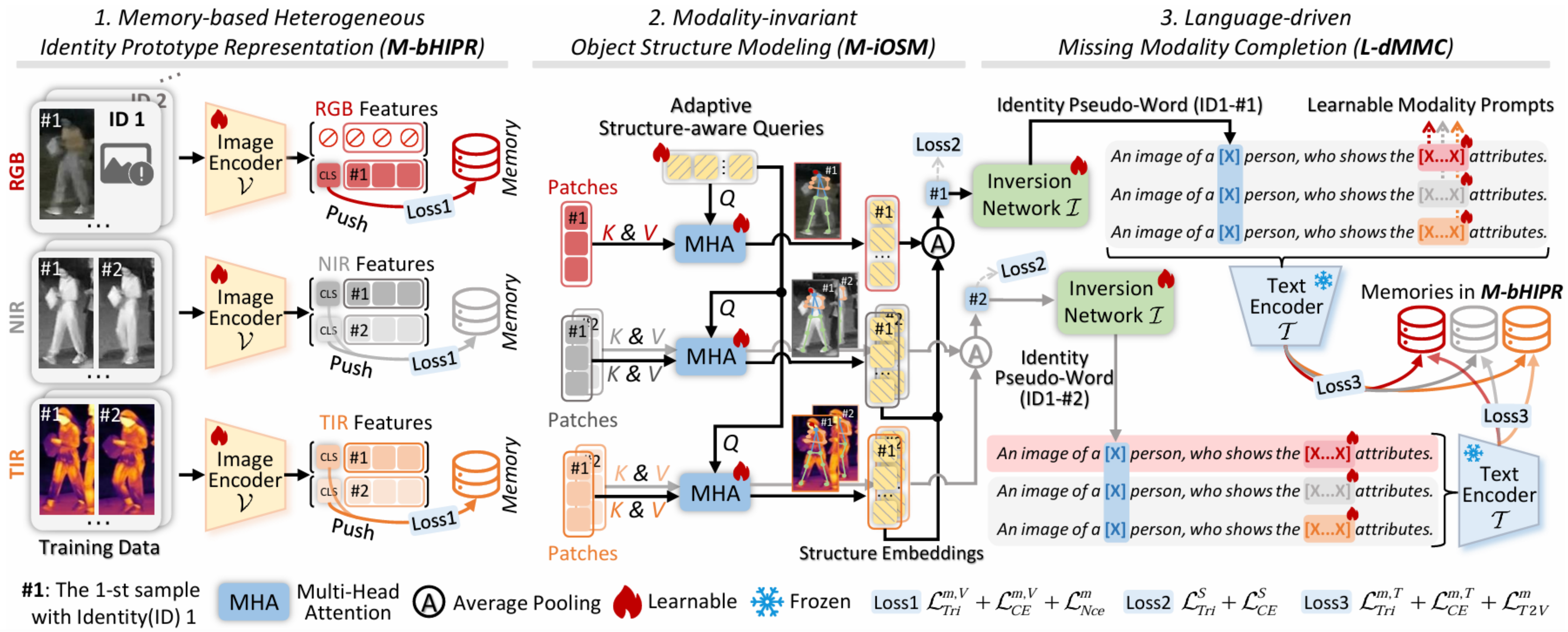
- Existing models trained on modality-complete datasets typically exhibit significantly degraded discrimination during inference with modality-incomplete inputs.

Motivation

- By harnessing VLMs' open-world vision-text alignment, text-derived semantic features may effectively compensate for incomplete visual information, enabling robust solutions for modality-missing training and inference.



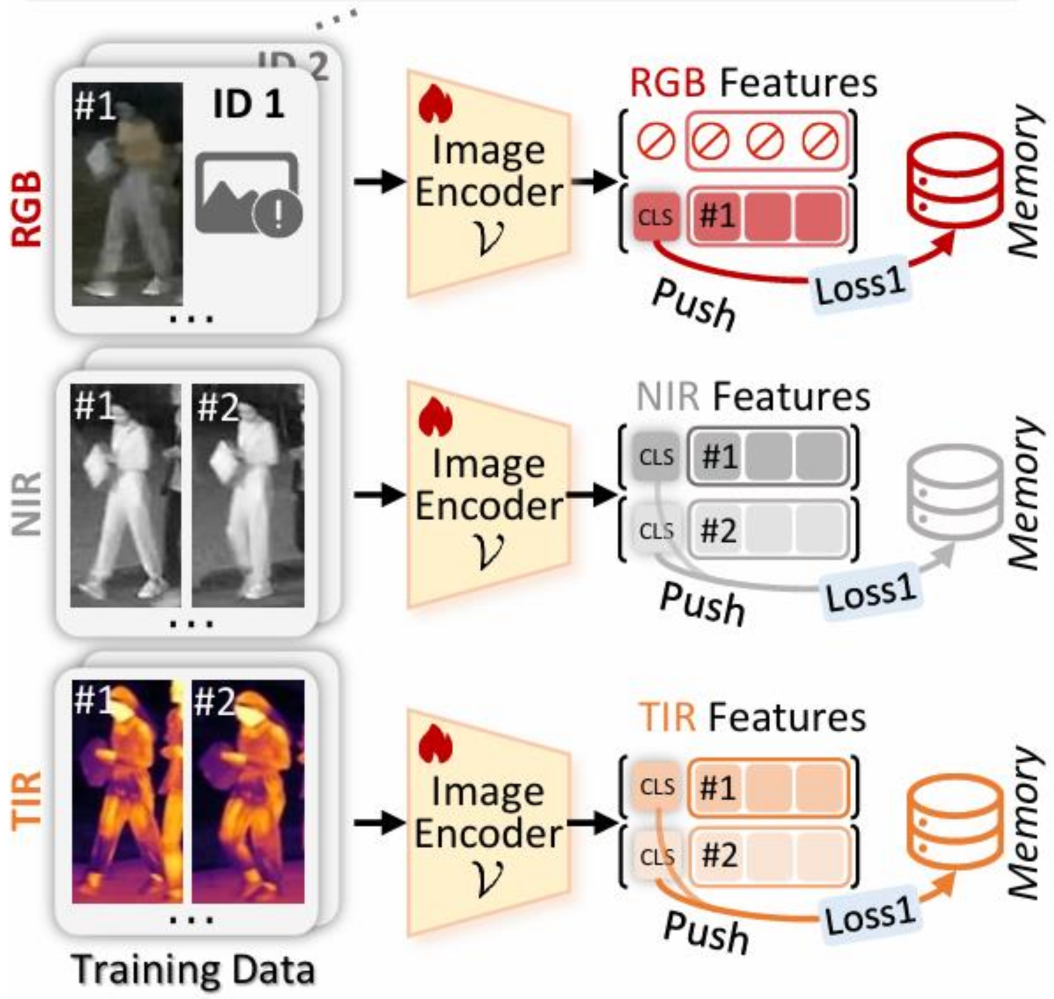
Framework



- Miss-ReID is the first work to handle multi-modality ReID under more general modality-missing scenarios encountered during both training and inference.

Methodology

1. M-bHIPR

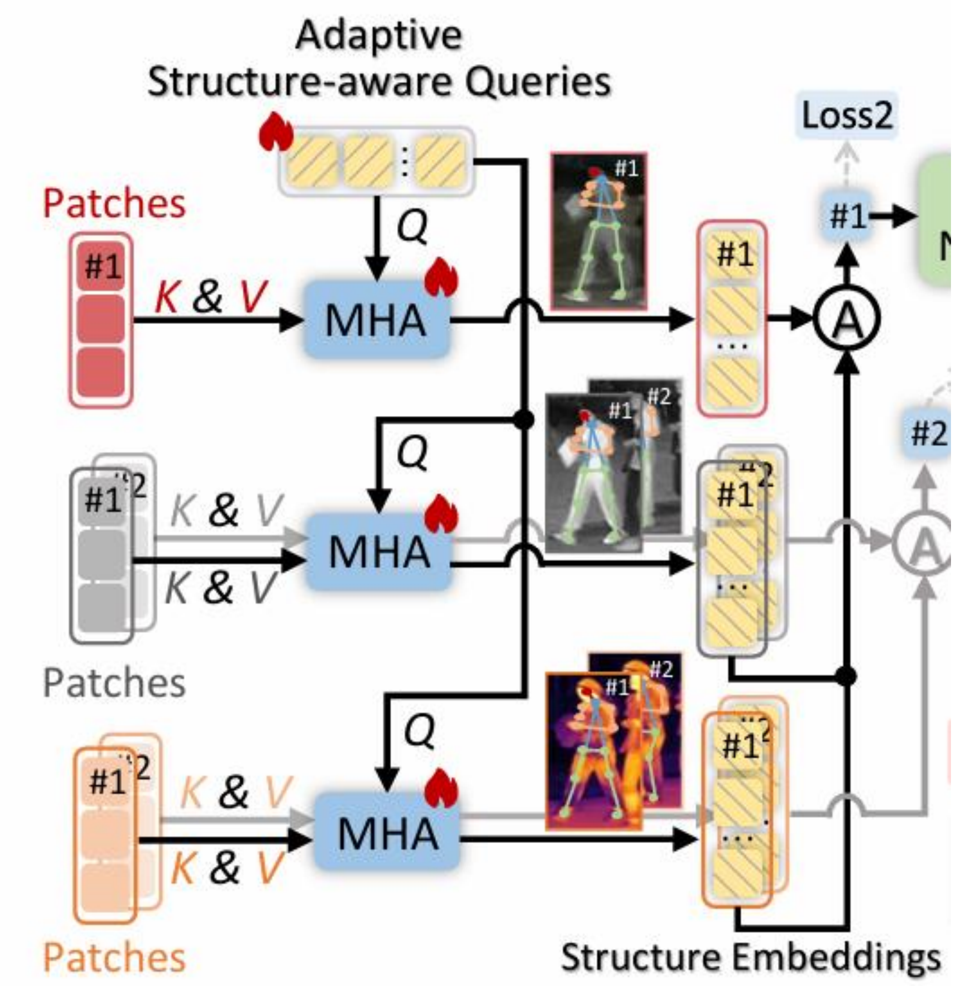


Memory-based Heterogeneous Identity Prototype Representation

Firstly, M-bHIPR extracts diverse visual features from accessible modalities, and then builds modality-specific memory banks to store heterogeneous prototypes for each individual identity, ensuring the preservation of multi-modality characteristics.

Methodology

2. M-iOSM

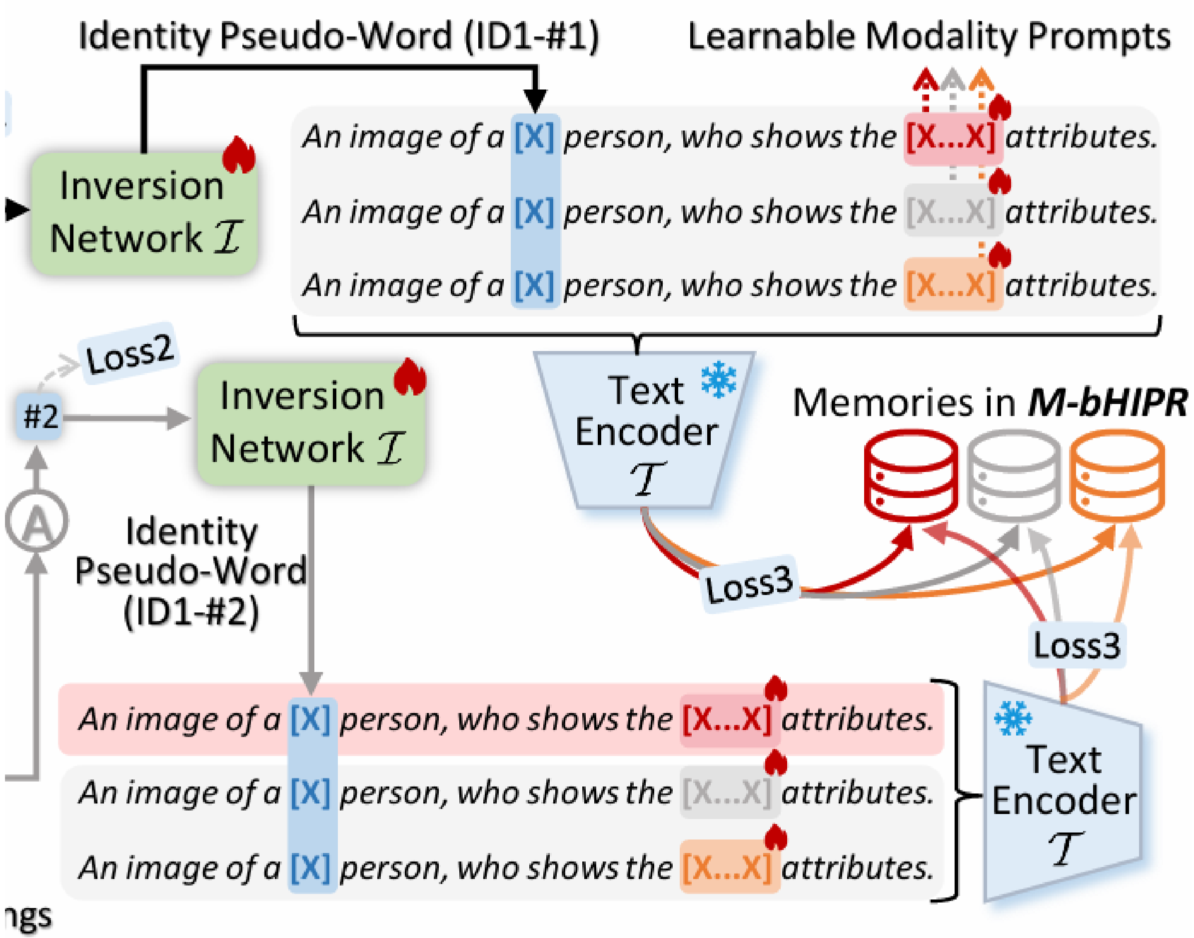


Modality-invariant
Object Structure Modeling

Afterwards, M-iOSM employs structure-aware query interactions to dynamically distill modality-invariant object structures from existing localized visual patches.

Methodology

3. L-dMMC



Language-driven Missing Modality Completion

- ✓ By leveraging the textual inversion technique, the extracted visual structural features are further reversed into pseudo-word tokens that encapsulate the identity-relevant structural semantics with L-dMMC module.
- ✓ Ultimately, the inverted tokens, integrated with diverse learnable modality prompts, are embedded into crafted textual templates to form the personalized linguistic descriptions for diverse modalities.
- ✓ Benefiting from VLMs' inherent vision-text alignment capability, L-dMMC produces the textual embeddings to substitute the absent visual cues.

Experiments

Ablation Study

Table 1: The impacts of various components. We report the comparison results between different combinations (Model **B** – **F**) and the baseline (Model **A**) under both **modality-complete** and **-missing** training settings on RGBNT201. Here, '**Modality Complete**' represents learning the modality-complete data during **training**, and '$\eta = (0.1, 0.1, 0.1)$' denotes randomly abandoning 10% RGB images, 10% NIR images, and 10% TIR images during **training**. The evaluations are both conducted across six modality-missing scenarios, and mean mAP and R-1 are reported below.

Index	Modules			Complexity		Modality Complete		$\eta = (0.1, 0.1, 0.1)$	
	M-bHIPR	M-iOSM	L-dMMC	**Params**	**FLOPs**	Mean **mAP**	Mean **R-1**	Mean **mAP**	Mean **R-1**
A	✗	✗	✗	86.4M	34.3G	48.9	50.4	46.4	47.0
B	✓	✗	✗	86.4M	34.3G	51.1(+2.2)	51.4(+1.0)	47.4(+1.0)	48.0(+1.0)
C	✗	✓	✗	86.4M	34.3G	50.2(+1.3)	52.1(+1.7)	46.9(+0.5)	48.7(+1.7)
D	✓	✓	✗	86.4M	34.3G	53.3(+4.4)	54.1(+3.7)	49.4(+3.0)	49.8(+2.8)
E	✗	✓	✓	89.6M	43.6G	52.1(+3.2)	53.4(+3.0)	47.4(+1.0)	49.7(+2.7)
F	✓	✓	✓	89.6M	43.6G	54.6(+5.7)	55.7(+5.3)	50.1(+3.7)	51.3(+4.3)

Experiments

Comparison with Stat-Of-The-Art Methods

NEURAL INFORMATION PROCESSING SYSTEMS

Table 2: Performance comparisons under **modality-missing** situations that only occur at the **inference** phase of multi-modality person ReID on RGBNT201. † denotes the model that is trained using both images and their corresponding text annotations. The best results are labeled with **boldface**. ↓x.x% and ↓x.x% highlight the lowest mAP and R-1 drop rates, respectively. '–' indicates that the metric is unpublished.

Methods	RNT		RNT		RNT		RNT		RNT		RNT		RNT		Mean	
	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1	mAP	R-1
PCB [ECCV 2018]	32.8	28.1	23.6 ↓28.0%	24.2 ↓13.9%	24.4 ↓25.6%	25.1 ↓10.7%	19.9 ↓39.3%	14.7 ↓47.7%	20.6 ↓37.2%	23.6 ↓16.0%	11.0 ↓66.5%	6.8 ↓75.8%	18.6 ↓43.3%	14.4 ↓48.8 %	19.7 ↓39.9%	18.1 ↓35.6%
TOP-ReID [AAAI 2024]	72.3	76.6	54.4 ↓24.8%	57.5 ↓24.9%	64.3 ↓11.1%	67.6 ↓11.7%	51.9 ↓28.2%	54.5 ↓28.9%	35.3 ↓51.2%	35.4 ↓53.8%	26.2 ↓63.8%	26.0 ↓66.1%	34.1 ↓52.8%	31.7 ↓58.6%	44.4 ↓38.6%	45.4 ↓40.7%
DeMo [AAAI 2025]	79.0	82.3	63.3 ↓19.9%	65.3 ↓20.7%	**72.6** ↓8.1%	**75.7** ↓8.0%	56.2 ↓28.9%	54.1 ↓34.3%	45.6 ↓42.3%	46.5 ↓43.5%	26.3 ↓66.7%	24.9 ↓69.7%	40.3 ↓49.0%	38.5 ↓53.2%	50.7 ↓35.8%	50.8 ↓38.3%
IDEA† [CVPR 2025]	**80.2**	**82.1**	62.9 ↓21.6%	– ↓–%	71.5 ↓10.8%	– ↓–%	58.4 ↓27.2%	– ↓–%	43.3 ↓46.0%	– ↓–%	27.1 ↓66.2%	– ↓–%	39.9 ↓50.2%	– ↓–%	50.5 ↓37.0%	– ↓–%
Miss-ReID [Ours]	76.9	78.9	**66.6** ↓13.4%	**68.2** ↓13.6%	72.4 ↓5.9%	75.5 ↓4.3%	**63.2** ↓17.8%	**63.8** ↓19.1%	**47.2** ↓38.6%	**49.5** ↓37.3%	**34.5** ↓55.1%	**33.3** ↓57.8%	**43.9** ↓42.9%	**44.3** ↓43.9%	**54.6** ↓29.0%	**55.7** ↓29.4%

Experiments

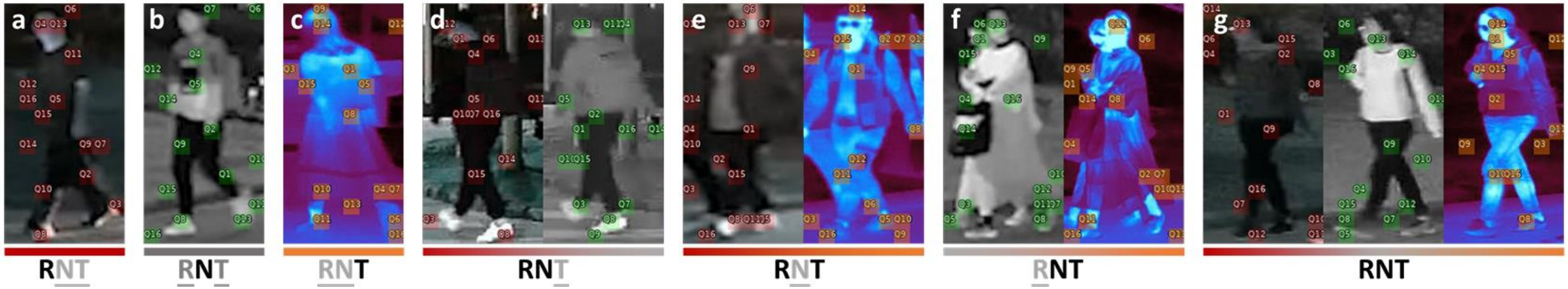
Performance Analysis of Miss-ReID under Varying Tri-modality Missing Rates

Table 3: Performance comparisons of setting different **tri-modality missing rates** on RGBNT201. Each tuple $(\eta_{rgb}, \eta_{nir}, \eta_{tir})$ represents the proportion of randomly abandoned RGB, Near-Infrared, and Thermal-Infrared images during **training**.

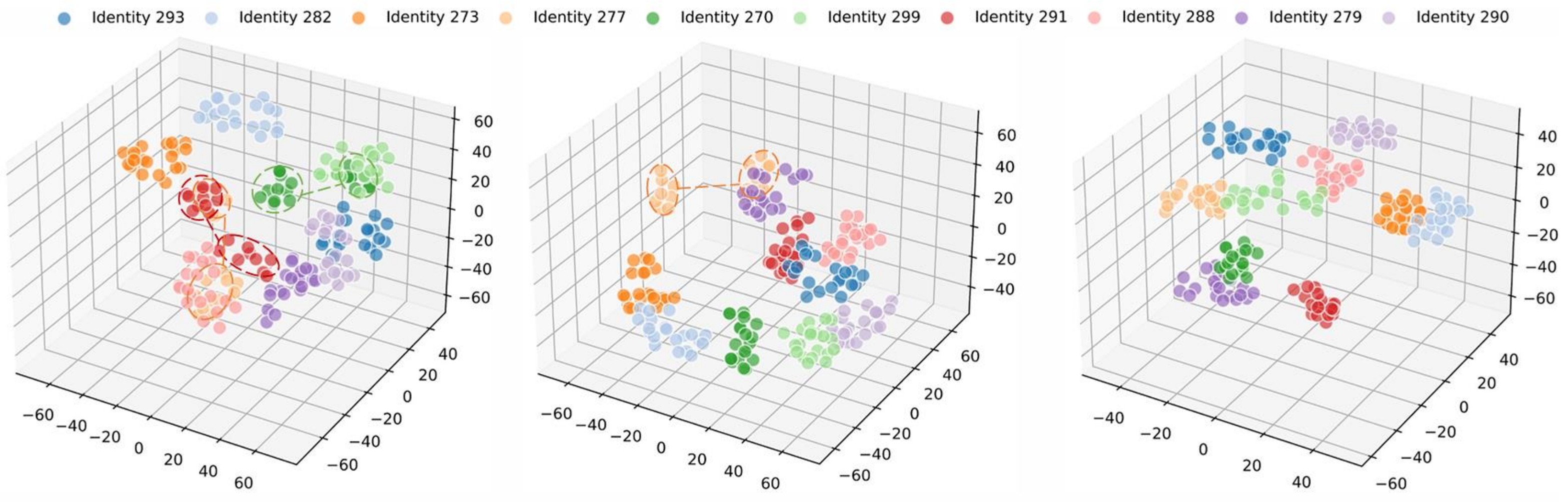
Tri-Modality Missing Rate η	(0.0, 0.0, 0.0)		(0.1, 0.1, 0.1)		(0.3, 0.3, 0.3)		(0.5, 0.5, 0.5)		(0.1, 0.3, 0.5)		(0.5, 0.3, 0.1)	
	mAP	**R-1**	**mAP**	**R-1**	**mAP**	**R-1**	**mAP**	**R-1**	**mAP**	**R-1**	**mAP**	**R-1**
RNT	76.9	78.9	72.3	73.4	68.4	71.2	68.2	72.8	69.6	72.2	67.6	67.6
RNT	66.6	68.2	61.3	61.7	57.6	58.3	56.7	58.4	56.1	58.4	57.6	58.4
RNT	72.4	75.5	68.8	72.6	65.8	69.5	63.6	65.1	66.9	69.7	65.7	67.0
RNT	63.2	63.8	55.3	55.7	52.3	56.5	52.3	54.4	53.2	57.3	50.2	50.7
RNT	47.2	49.5	42.8	45.2	44.9	47.5	41.6	40.8	41.1	40.3	47.1	47.6
RNT	34.5	33.3	30.9	29.7	26.8	26.3	26.5	22.2	27.1	28.1	26.5	24.6
RNT	43.9	44.3	41.5	42.7	43.0	45.5	42.7	46.4	43.9	47.0	39.2	38.8
Mean	**54.6**	**55.7**	**50.1**	**51.3**	**48.4**	**50.6**	**47.3**	**47.9**	**48.1**	**50.1**	**47.7**	**47.8**

Experiments

Visualizations



Visualizations of the attentive regions towards 16 well-learned structure-aware queries.



(a) Only Visual Features (NIR) (b) Concatenating RGB Textual Features (c) Concatenating TIR Textual Features

The feature distributions of 10 random identities.

Experiments

Retrieval Result

(a) Baseline

(b) Miss-ReID

Thanks for watching!

Ruida Xi