SAFR: FOUNDATION-ERROR-AWARE RANKING AND SPARSE AFFINITY RESIDUALIZATION FOR MULTI-MODAL OBJECT RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modal object Re-Identification (ReID) combines RGB, near-infrared, and thermal-infrared observations to match identities across cameras or scenes. A frozen visual encoder can rank a look-alike distractor above the true match when the latter is seen from a different viewpoint. We propose Sparse Foundation-Affinity Residualization (SAFR), which uses frozen-model reversals to shape training margins and frozen affinities to correct retrieval scores. During training, identity labels reveal such reversals, and foundation-error-aware ranking increases the task margin in proportion to their magnitude while retaining standard identity supervision. After training, sparse affinity correction estimates, on different-identity training pairs, the part of the task score associated with global and local frozen affinities. Held-out training identities determine which contributions to subtract, yielding a fixed retrieval score. Experiments obtain 83.7%, 89.6%, and 71.5% mAP on RGBNT201, RGBNT100, and MSVR310, respectively. Component and mechanism analyses further show how representation learning and score correction contribute to the final retrieval performance.

1 INTRODUCTION

Object Re-Identification retrieves the same person or vehicle across cameras and scenes. Combining synchronized RGB, near-infrared (NIR), and thermal-infrared (TIR) observations provides complementary appearance cues when visible-light images are degraded. Effective retrieval nevertheless requires more than visual similarity: the correct identity must outrank candidates that share clothing, vehicle shape, or background but belong to different targets.

Multi-modal ReID methods have developed from modality-specific feature learning and fusion (Zheng et al., 2021; Li et al., 2020; Wang et al., 2022; Crawford et al., 2023) toward token interaction, prompt adaptation, and uncertainty-aware integration (Wang et al., 2024a; Zhang et al., 2024; Lin et al., 2025; Zhang et al., 2025; Wan et al., 2025). Large-scale visual pretraining, including DINOv3 (Siméoni et al., 2025), further provides transferable global and local features. These advances improve the information available to a retrieval model. This raises a question: when pretrained visual relations disagree with identity labels, how should they shape the learned representation and matching score?

Our starting point is a concrete ordering error: for a query vehicle, a different-identity car with similar color and shape may receive a higher frozen-encoder similarity than the true vehicle seen from another camera. Identity labels expose this reversal. Task adaptation can still retain affinity for the look-alike distractor, so we use the frozen encoder in two ways. On task-mined hard pairs, a frozen-reference reversal enlarges the task ranking margin; separately, reference affinities associated with variation in task scores among different-identity pairs are candidates for subtraction at retrieval. Fitting only on such pairs focuses the correction on nonmatches without fitting the high scores of true matches. Held-out training identities decide which reference contributions to retain. SAFR implements these steps with an identity-trained task encoder and a frozen visual reference.

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.

During training, *foundation-error-aware ranking* operates on hard positive-negative pairs selected by the task encoder. A frozen-reference reversal increases the required task margin in proportion to its magnitude. When the reference order is correct, ordinary smooth margin ranking remains active. Classification and triplet losses continue to supervise identity discrimination. This construction makes the reference error an additional training signal without using its ordering as a target.

After training, *sparse foundation-affinity correction* refines the matching score with both encoders fixed. It combines modality-wise task similarities and estimates subtractive global and local reference contributions on different-identity training pairs. A reference subset is selected using held-out training identities, and the resulting coefficients remain fixed during retrieval. Identity supervision thus determines both the strength and the selection of the reference contributions.

Experiments on RGBNT201, RGBNT100, and MSVR310 examine these two uses of the frozen reference. The results show consistent improvements on person and vehicle benchmarks, while component and mechanism analyses distinguish the effects of representation learning, modality weighting, and reference correction. Our contributions are:

- A foundation-error-aware training objective that increases an identity-ranking margin according to a frozen reference's reversal on task-mined pairs.
- A sparse affinity-correction procedure that combines negative-pair coefficient estimation with reference-subset selection on held-out training identities.
- Extensive experiments on three multi-modal ReID benchmarks, together with component, mechanism, reference, and efficiency analyses.

2 Related Work

2.1 Multi-modal Object Re-Identification

Multi-modal Object ReID exploits complementary information from RGB, near-infrared, and thermal-infrared observations to preserve identity discrimination under changing illumination and imaging conditions. PFNet, HAMNet, and Interact–Embed–EnlargE learn complementary features through progressive fusion or cross-modal interaction (Zheng et al., 2021; Li et al., 2020; Wang et al., 2022). GraFT uses learnable fusion tokens (Yin et al., 2023), whereas UniCat concatenates features from independently trained unimodal encoders (Crawford et al., 2023). Finer-grained interaction uses token permutation, token selection, or selective spectral coupling (Wang et al., 2024a; Zhang et al., 2024; Yu et al., 2025); prompt-based methods also coordinate spectral information (Lin et al., 2025; Zhang et al., 2025). STMI (Xu et al., 2026) uses segmentation-guided feature modulation and cross-modal hypergraph interaction. HTT uses test-time training to address target-domain shift (Wang et al., 2024b); FACENet targets flare-degraded observations (Zheng et al., 2025); DESANet handles low-quality or missing modalities with enhancement, reconstruction, and soft alignment (Dong et al., 2025); and UGG-ReID models uncertainty in noisy local cues (Wan et al., 2025). MFRNet and DeMo coordinate modality-shared and modality-specific features with expert modules (Feng et al., 2025b; Wang et al., 2025b). These approaches strengthen identity descriptors through complementary spectra; SAFR examines errors in pairwise affinities supplied by a separate frozen encoder.

2.2 Pretrained Priors and Knowledge Transfer

Pretrained models can supply multi-modal ReID with visual representations, textual semantics, or teacher knowledge. TransReID applies vision Transformers to identity learning (He et al., 2021), while DINOv3 provides transferable visual features (Siméoni et al., 2025). CLIP's image-text alignment (Radford et al., 2021) motivates identity-related prompts in CLIP-ReID and ICPL-ReID (Li et al., 2023; 2026c). DMPT adapts a pretrained visual backbone with modality-aware prompts (Lin et al., 2025), whereas AIO uses a frozen shared encoder for RGB, infrared, sketch, and text retrieval (Li et al., 2024). IDEA (Wang et al., 2025c) uses generated modality-specific descriptions as semantic guidance, while PEFT-BoA (Li et al., 2026a) adapts a pretrained CLIP visual encoder through adapters. Relational knowledge distillation transfers teacher-defined sample relations (Park et al., 2019); PMKD (Zheng et al., 2026a) transfers multi-modal teacher knowledge to modality-specific models using sample-hardness guidance. SAFR keeps a separate visual encoder frozen

and uses its affinities as pairwise references; identity labels determine how those references shape task-ranking constraints and retrieval-score correction.

2.3 Ranking Learning and Retrieval-Score Refinement

Beyond representations, the relative order of positive and negative candidates also affects ReID. Batch-hard triplet learning (Hermans et al., 2017) constrains difficult pairs, while FastAP optimizes an approximation to ranked-list average precision (Cakir et al., 2019). After feature extraction, k-reciprocal re-ranking uses gallery neighborhoods to revise initial distances (Zhong et al., 2017), while metric-fusion re-ranking combines multiple similarity relations (Bai et al., 2019). SAFR uses frozen-model affinity as a common reference: its identity-ordering errors adjust the training ranking constraint, and training identities determine a fixed pairwise retrieval score. Thus, training-time ranking and retrieval-time correction address the same type of affinity error.

3 Method

For target i, let

$$\mathcal{I}_i = \{I_i^m \mid m \in \mathcal{M}\}, \qquad \mathcal{M} = \{R, N, T\}, \tag{1}$$

denote its aligned RGB, NIR, and TIR observations, with identity label y_i. SAFR learns retrieval features with a task encoder f_θ and obtains reference affinities from a frozen foundation encoder f_0. Figure 1 summarizes the training and retrieval stages.

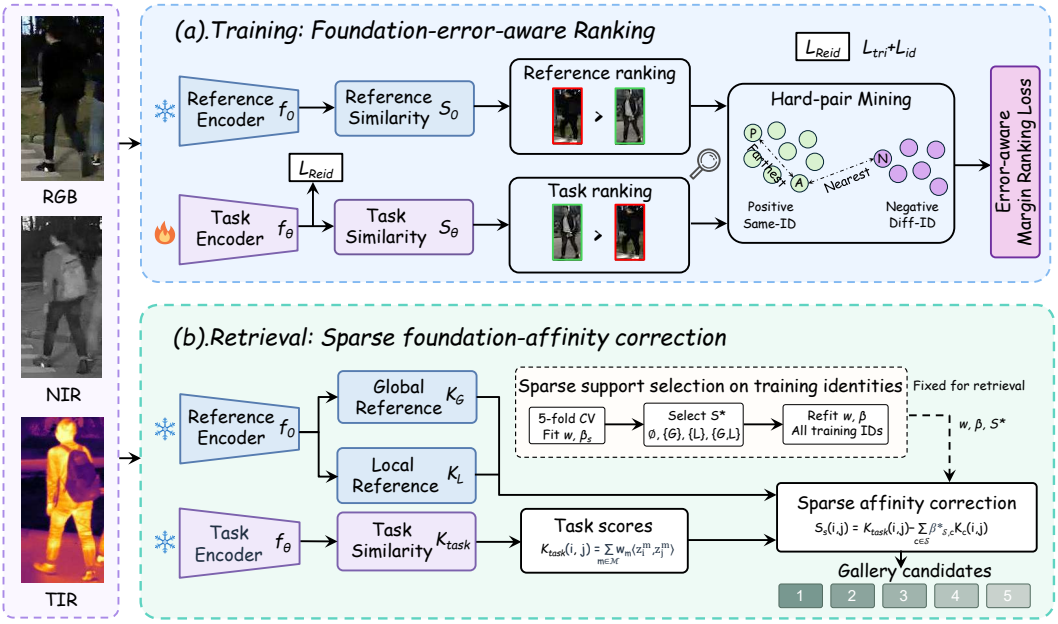

Figure 1: Overview of SAFR. (a) During training, a frozen reference exposes identity-ordering reversals on task-mined hard pairs, increasing the task encoder's ranking margin. (b) After training, both encoders are fixed. Training identities determine modality weights, global and local reference coefficients, and the selected reference support, yielding a fixed retrieval score.

3.1 Foundation-Error-Aware Ranking

For modality m, f_θ^m denotes the shared backbone with its modality-specific adapter. It extracts a normalized feature

$$z_i^m = \mathrm{Norm}(f_\theta^m(I_i^m)). \tag{2}$$

Here, Norm denotes ℓ_2 normalization. Using $[\,;\,]$ for concatenation gives

$$h_i = \mathrm{Norm}\big([z_i^R; z_i^N; z_i^T]\big), \tag{3}$$

which is used by identity classification and triplet learning. With $\langle \cdot, \cdot \rangle$ denoting the Euclidean inner product, task similarity is

$$s_\theta(i,j) = \langle h_i, h_j \rangle = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \langle z_i^m, z_j^m \rangle. \tag{4}$$

Frozen foundation reference. Let u_i^m be the normalized local feature extracted by the frozen encoder for modality m; its construction is specified in Section 3.2. The reference similarity is

$$s_0(i,j) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \langle u_i^m, u_j^m \rangle. \tag{5}$$

The reference may rank a different-identity candidate above a true match. Identity labels let us detect this ordering error during training rather than require the task encoder to imitate the reference.

Error-conditioned ranking margin. For anchor i in a mini-batch of size B, let $\mathcal{P}_i^B$ contain same-identity samples from other camera/scene groups, falling back to other same-identity samples if none exist. Let $\mathcal{N}_i^B$ contain different-identity samples. Identity-balanced sampling ensures both sets are nonempty. Hard examples are selected by task similarity alone:

$$p_i = \arg\min_{j \in \mathcal{P}_i^B} s_\theta(i,j), \qquad n_i = \arg\max_{j \in \mathcal{N}_i^B} s_\theta(i,j). \tag{6}$$

At these indices, the frozen reference's positive-negative reversal and the resulting task margin are

$$c_i = [\mathrm{sg}(s_0(i,n_i) - s_0(i,p_i))]_+, \qquad d_i = \delta + \alpha c_i, \tag{7}$$

where sg stops gradient propagation, $[x]_+ = \max(x, 0)$, $\delta \geq 0$ is the ordinary margin, and $\alpha \geq 0$ controls the increment. We optimize

$$\mathcal{L}_{\mathrm{rank}} = \frac{1}{B} \sum_{i=1}^{B} \tau \log \left[1 + \exp\left(\frac{s_\theta(i,n_i) - s_\theta(i,p_i) + d_i}{\tau} \right) \right], \tag{8}$$

where $\tau > 0$ controls smoothing. The loss applies to all anchors. When the frozen reference does not place the selected negative above the positive, $c_i = 0$ and the ordinary smooth margin remains active. A reference reversal gives $c_i > 0$ and requires a larger task margin, encouraging the task encoder to separate the identities despite that error. Gradients pass through task scores only; standard ReID losses are described in Section 3.3. At $\alpha = 0$, Eq. (8) is the ordinary smooth ranking loss.

3.2 Sparse Foundation-Affinity Correction

The ranking loss shapes the task representation but does not determine the final retrieval score. After training, we fix both encoders, learn modality weights, and estimate how foundation affinities co-vary with task scores on different-identity training pairs. Using only nonmatches prevents the high scores of true matches from determining the reference component to subtract. Held-out training identities select which references, if any, to subtract. The fitted association is statistical, not a causal decomposition of appearance and identity.

Foundation-affinity dictionary. For reference type $c \in \{G, L\}$, let $v_{i,c}^m$ be a normalized frozen feature. The global reference G uses the class token. For the local reference L, we ℓ_2-normalize patch tokens, average-pool them over a 2×2 spatial grid, normalize the four pooled vectors, and normalize their concatenation. The pairwise affinity is

$$K_c(i,j) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \langle v_{i,c}^m, v_{j,c}^m \rangle. \tag{9}$$

The dictionary $\mathcal{A} = \{K_G, K_L\}$ thus contains global and local affinities; $u_i^m = v_{i,L}^m$ in Eq. (5). We combine modality-wise task similarities as

$$K_{\mathrm{task}}(i,j) = \sum_{m \in \mathcal{M}} w_m \langle z_i^m, z_j^m \rangle, \quad w_m \geq 0, \quad \sum_m w_m = 1. \tag{10}$$

We fit non-negative modality coefficients and an intercept by ridge regression on sampled training pairs with binary same-identity targets, assigning half of the total weight to each pair class. Within every fitting split, inner five-fold identity cross-validation selects the ridge strength by mean class-balanced squared error. We refit and normalize the modality coefficients to obtain $\mathbf{w}$; the intercept is omitted because it does not affect candidate ranking.

Reference-based coefficient estimation. For a candidate support $\mathcal{S} \subseteq \{G, L\}$, we regress task scores on foundation affinities using different-identity training pairs. Define

$$\mathcal{N}_i = \{j \mid y_j \neq y_i\}, \qquad N_i = |\mathcal{N}_i|, \tag{11}$$

and let $\mathbf{k}^i_{\text{task}} \in \mathbb{R}^{N_i}$ contain task similarities to these samples. The columns of $X_i^{\mathcal{S}} \in \mathbb{R}^{N_i \times |\mathcal{S}|}$ contain their foundation affinities. We center each query with

$$H_i = I_{N_i} - \frac{1}{N_i} \mathbf{1}_{N_i} \mathbf{1}_{N_i}^{\top}. \tag{12}$$

Here, I_{N_i} is the identity matrix and $\mathbf{1}_{N_i}$ is the all-ones vector. We estimate bounded non-negative coefficients by

$$\boldsymbol{\beta}^*_{\mathcal{S}} = \arg \min_{0 \leq \boldsymbol{\beta} \leq \beta_{\max} \mathbf{1}} \sum_i \left\| H_i \mathbf{k}^i_{\text{task}} - H_i X_i^{\mathcal{S}} \boldsymbol{\beta} \right\|_2^2. \tag{13}$$

The sum runs over anchors in the fitting split, $\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{S}|}$, and $\beta_{\max}$ is a shared coefficient bound. For a fixed query, candidate rankings depend on score differences rather than a shared offset. Centering removes that offset from both task scores and reference affinities, so the coefficients reflect their covariation across the query's nonmatches. The non-negative coefficients yield subtractive reference terms in the candidate score:

$$S_{\mathcal{S}}(i, j) = K_{\text{task}}(i, j) - \sum_{c \in \mathcal{S}} \beta^*_{\mathcal{S},c} K_c(i, j). \tag{14}$$

The subtraction applies to every candidate, including true matches. We therefore select reference terms by held-out positive-negative ranking gaps rather than by the nonmatch regression fit alone.

Sparse support selection on training identities. We evaluate four supports: $\varnothing$, $\{G\}$, $\{L\}$, and $\{G, L\}$. Here, sparsity refers only to selecting a subset of the two reference functions. The empty support retains K_{task}; disabling the entire correction stage also fixes $w_m = 1/|\mathcal{M}|$. In outer five-fold identity cross-validation, each fold fits $\mathbf{w}$ and $\boldsymbol{\beta}_{\mathcal{S}}$ on four folds and uses the fifth-fold identities as queries against all training samples. Held-out query identities never enter parameter fitting, although fitting-fold identities may appear as negative gallery candidates. Let $\mu_i^{\mathcal{S}}$ and $\sigma_i^{\mathcal{S}}$ be the mean and standard deviation of query i's different-identity gallery scores. We compute

$$\widehat{S}_{\mathcal{S}}(i, j) = \frac{S_{\mathcal{S}}(i, j) - \mu_i^{\mathcal{S}}}{\max(\sigma_i^{\mathcal{S}}, \epsilon)}, \tag{15}$$

where $\epsilon > 0$ prevents division by zero. This selection-only standardization preserves within-query rankings while making gaps comparable across queries. Here N_i counts different-identity gallery samples; $\mathcal{P}_i$ contains same-identity samples from other camera/scene groups, and $\mathcal{H}_i^k(\mathcal{S})$ contains the $k_i = \min(k, N_i)$ highest-scoring different-identity samples, without requiring distinct identities. For queries with nonempty positive and negative sets, the hard ranking gap is

$$\gamma_i(\mathcal{S}) = \min_{p \in \mathcal{P}_i} \widehat{S}_{\mathcal{S}}(i, p) - \frac{1}{k_i} \sum_{n \in \mathcal{H}_i^k(\mathcal{S})} \widehat{S}_{\mathcal{S}}(i, n). \tag{16}$$

We select $\mathcal{S}^*$ by the mean of the five fold-average gaps, favoring fewer references when scores differ by at most 10^{-6}. After refitting $\mathbf{w}$ and $\boldsymbol{\beta}$ on all training identities, we use the unstandardized test-time score

$$S(i, j) = S_{\mathcal{S}^*}(i, j). \tag{17}$$

3.3 Optimization and Retrieval

Let $\mathcal{H} = \{h_i\}_{i=1}^B$ be the batch of descriptors. We supervise the task encoder with label-smoothed cross-entropy and batch-hard triplet losses:

$$\mathcal{L}_{\text{ReID}} = \mathcal{L}_{\text{id}}(\mathcal{H}) + \mathcal{L}_{\text{tri}}(\mathcal{H}). \tag{18}$$

For the first five epochs, we optimize only $\mathcal{L}_{\text{ReID}}$. From epoch 6 to epoch 20, we add the error-conditioned ranking loss in Eq. (8), jointly optimizing all three terms:

$$\mathcal{L} = \mathcal{L}_{\text{ReID}} + \mathcal{L}_{\text{rank}}. \tag{19}$$

Only the task encoder is updated; the foundation encoder remains frozen throughout training. Afterward, we fit the modality weights, reference subset, and correction coefficients on training identities as described in Section 3.2. At retrieval, both encoders extract features, and Eq. (17) scores and ranks the gallery candidates for each query.

4 EXPERIMENTS

4.1 DATASETS AND EVALUATION PROTOCOL

We evaluate SAFR on three public benchmarks. RGBNT201 (Zheng et al., 2021) contains 201 person identities and 4,787 aligned RGB-NIR-TIR triplets with substantial pose, viewpoint, background, and illumination variation. RGBNT100 (Li et al., 2020) contains 100 vehicle identities and 17,250 triplets; vehicles of the same model and color require fine-grained ranking. MSVR310 (Zheng et al., 2023) contains 310 vehicle identities and 2,087 triplets, with cross-environment observations, long temporal gaps, and fewer samples per identity. We follow each dataset's official evaluation protocol and report mean average precision (mAP) and Rank-1/5/10 accuracy.

4.2 IMPLEMENTATION DETAILS

Both the task encoder and frozen reference encoder use DINOv3 ViT-B/16 (Siméoni et al., 2025). Within each encoder, the three modalities share a backbone, and the task encoder uses modality-specific adapters. Global and local references use the block-8 class token and block-10 patch tokens, respectively. Input resolution is 256×128 for RGBNT201 and 128×256 for RGBNT100 and MSVR310. Data augmentation includes random horizontal flipping, cropping, and erasing. We train with AdamW for 20 epochs, using eight identities and eight images per identity in each batch. The learning rates for the backbone and other modules are 3×10^{-5} and 3.5×10^{-4}, respectively, with weight decay 0.05 and a five-epoch learning-rate warm-up. The first five epochs use classification and triplet losses; from epoch 6 onward, the ranking loss is optimized jointly with them. We use $\alpha = 1$, $\delta = 0.2$, $\tau = 0.12$, $k = 5$, and $\beta_{\max} = 3$ by default.

4.3 COMPARISON WITH STATE OF THE ART

Table 1 compares methods without task-specific text or segmentation masks. Symbols indicate backbone families: $*$ denotes other ViTs, † CLIP, and ‡ DINOv3; unmarked methods use CNNs. The ProxyTTT‡ and PEFT-BoA‡ results are our DINOv3-B adaptations of public implementations.

Table 1: Comparison on RGBNT201, RGBNT100, and MSVR310 without task-specific external priors (%). Bold and underlined values indicate the best and second-best results for each metric, respectively; ties share the same formatting. Both ProxyTTT rows report results with test-time training disabled (w/o TTT). Dashes indicate unreported results.

Method	RGBNT201				RGBNT100		MSVR310	
	mAP	R-1	R-5	R-10	mAP	R-1	mAP	R-1
HAMNet (Li et al., 2020)	27.7	26.3	41.5	51.7	74.5	93.3	27.1	42.3
PFNet (Zheng et al., 2021)	38.5	38.9	52.0	58.4	68.1	94.1	23.5	37.4
CCNet (Zheng et al., 2023)	–	–	–	–	77.2	96.3	36.4	55.2
LRMM (Wu et al., 2025)	52.3	53.4	64.6	73.2	78.6	96.7	36.7	49.7
UniCat* (Crawford et al., 2023)	57.0	55.7	–	–	79.4	96.2	–	–
TOP-ReID* (Wang et al., 2024a)	72.3	76.6	84.7	89.4	81.2	96.4	35.9	44.6
EDITOR* (Zhang et al., 2024)	66.5	68.3	81.1	88.2	82.1	96.4	39.0	49.3
MambaPro† (Wang et al., 2025a)	78.9	83.4	89.8	91.9	83.9	94.7	47.0	56.5
DeMo† (Wang et al., 2025b)	79.0	82.3	88.8	92.0	86.2	97.6	49.2	59.8
MFRNet† (Feng et al., 2025b)	80.7	83.6	91.9	94.1	88.2	97.4	50.6	64.8
Signal† (Liu et al., 2026)	80.3	85.2	91.4	93.7	86.3	97.6	53.6	71.9
FUSE† (Qi et al., 2026)	81.4	<u>86.1</u>	91.5	93.8	<u>88.5</u>	96.9	50.1	65.7
MDReID† (Feng et al., 2025a)	82.1	85.2	90.3	92.6	85.3	95.6	51.0	68.9
ProxyTTT† (Zheng et al., 2026b)	82.3	84.7	90.6	92.7	88.4	97.9	62.1	71.7
ProxyTTT‡	76.7	78.9	87.4	90.6	87.0	97.6	<u>63.6</u>	79.2
PEFT-BoA† (Li et al., 2026a)	<u>82.7</u>	<u>86.1</u>	<u>92.3</u>	<u>94.7</u>	85.1	97.8	57.4	74.5
PEFT-BoA‡	79.8	82.4	88.9	91.0	86.1	<u>98.3</u>	63.0	<u>82.7</u>
SAFR‡ (ours)	**83.7**	**86.7**	**93.5**	**95.7**	**89.6**	**99.5**	**71.5**	**87.1**

Multi-modal person ReID. On RGBNT201, SAFR obtains 83.7% mAP and 86.7% Rank-1. Relative to the ViT-based TOP-ReID, it gains 11.4 and 10.1 points on these metrics. Among CLIP-based methods, SAFR exceeds MFRNet by 3.0 mAP and 3.1 Rank-1 points, and PEFT-BoA†—the strongest competitor in mAP and tied for the strongest competing Rank-1—by 1.0 and 0.6 points. Its Rank-5 and Rank-10 also exceed PEFT-BoA† by 1.2 and 1.0 points. Under the same DINOv3-B backbone, SAFR surpasses our PEFT-BoA‡ adaptation by 3.9 mAP and 4.3 Rank-1 points.

Multi-modal vehicle ReID. On RGBNT100, SAFR achieves 89.6% mAP and 99.5% Rank-1. It improves upon MFRNet by 1.4 and 2.1 points, respectively. It also exceeds the strongest competing mAP from FUSE (88.5%) by 1.1 points and the strongest competing Rank-1 from PEFT-BoA‡ (98.3%) by 1.2 points. On MSVR310, SAFR reaches 71.5% mAP and 87.1% Rank-1, gaining 17.9 and 15.2 points over Signal. Against the strongest competing mAP from ProxyTTT‡ (63.6%) and Rank-1 from PEFT-BoA‡ (82.7%), the gains remain 7.9 and 4.4 points. SAFR therefore leads both vehicle benchmarks across methods with different visual backbones.

Comparison with task-specific external priors. Table 2 separately compares systems that use generated text or segmentation masks. Among text-based methods, IDEA obtains 80.2% mAP and 82.1% Rank-1, while NEXT reaches 82.4% and 86.6%. STMI, which uses both text and masks, obtains 81.2% and 83.4%. Without either external prior, SAFR reaches 83.7% mAP and 86.7% Rank-1, exceeding NEXT by 1.3 and 0.1 points and STMI by 2.5 and 3.3 points. With the same DINOv3-B visual-backbone architecture, it also exceeds the adapted IDEA by 2.4 mAP and 3.5 Rank-1 points. These full-system comparisons show that SAFR achieves strong retrieval without generated descriptions or segmentation masks; they do not isolate the effect of either prior.

Table 2: Comparison with methods using task-specific external priors on RGBNT201 (%).

Method	Visual backbone	Extra prior	mAP	Rank-1
IDEA (Wang et al., 2025c)	CLIP-B/16	Text	80.2	82.1
IDEA (DINOv3-B) (Gao et al., 2026)	DINOv3-B	Text	81.3	83.2
NEXT (Li et al., 2026b)	CLIP-B/16	Text	<u>82.4</u>	<u>86.6</u>
STMI (Xu et al., 2026)	CLIP-B/16	Text + mask	81.2	83.4
SAFR (ours)	DINOv3-B	None	**83.7**	**86.7**

4.4 Ablation Studies

On RGBNT201, the baseline trains the task encoder with identity classification and triplet losses, then retrieves by equally averaging its RGB, NIR, and TIR similarities. The ranking branch adds the foundation-error-aware ranking loss from epoch 6 onward but retains this equal-weight retrieval score. For affinity correction, we fix the trained encoder, learn modality weights and correction coefficients, and select the reference support using training identities before retrieving with the corrected score. Correction alone starts from the baseline-trained encoder, whereas the full method uses the ranking-trained encoder.

Overall components. Table 3 compares the four configurations. Ranking alone raises mAP from 75.7% to 76.5% and Rank-1 from 76.0% to 78.6%. Correction alone obtains 82.3% mAP and 84.0% Rank-1. Combining both reaches 83.7% mAP and 86.7% Rank-1. Thus, the ranking objective still improves identity discrimination after the retrieval scores have been recalibrated.

Correction mechanism. Table 4 fixes the error-aware-trained task model, learned modality weights, and global-local references, changing only the final correction rule. The uncorrected weighted score obtains 78.2% mAP; unlike the uncorrected row in Table 3, it already uses learned modality weights. Fixed-coefficient subtraction with $\beta_G = \beta_L = 0.15$ reaches 79.6% mAP, whereas fitting coefficients without query centering gives 78.4%. Centered coefficient learning increases mAP to 83.7%, a gain of 5.5 points over the uncorrected weighted score and 5.3 points over uncentered fitting; Rank-1 rises from 79.9% to 86.7% relative to the uncorrected weighted score. By removing query-specific offsets before fitting different-identity pairs, centering directs the fit toward score variation among candidates

Table 3: Component ablation on RGBNT201 (%). Rows without affinity correction use equal modality weights; corrected rows also learn modality weights.

Foundation-error-aware ranking	Affinity correction	mAP	Rank-1	Rank-5	Rank-10
–	–	75.7	76.0	84.4	88.2
✓	–	76.5	78.6	85.2	88.9
–	✓	82.3	84.0	90.8	94.3
✓	✓	**83.7**	**86.7**	**93.5**	**95.7**

for the same query. The controlled comparison shows that learning subtraction coefficients alone does not account for the full improvement.

Table 4: Ablation of score-correction mechanisms on RGBNT201 (%).

Correction rule	mAP	Rank-1	Rank-5	Rank-10
Weighted task score, no subtraction	78.2	79.9	87.1	91.4
Fixed-coefficient subtraction	79.6	81.6	87.8	92.3
Learned coefficients, no query centering	78.4	80.3	87.1	91.5
Centered coefficient learning (full)	**83.7**	**86.7**	**93.5**	**95.7**

Figure 2 separates the intended score-level change from its effect on identity ranking. In panel (a), the Pearson correlation between retrieval scores and the fitted foundation-reference component shifts from 0.256 to -0.036 after subtraction. This association change follows the score construction; panel (b) shows the ranking effect. Among 836 queries, correction enlarges hard-ranking margins for 73.7%, and the mean margin increases from 0.015 to 0.175. It repairs 69 incorrect top matches; overall, the number of queries with a correct top match increases by 57, consistent with the Rank-1 increase in Table 4.

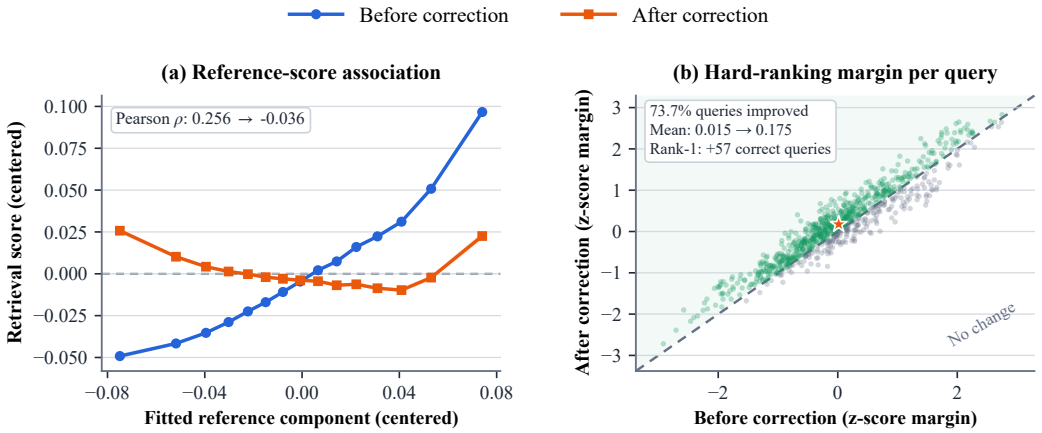

Figure 2: Mechanism analysis on RGBNT201. (a) Association between the fitted foundation-reference component and centered retrieval scores before and after correction on different-identity pairs; curves show equal-frequency bin means and shading denotes the standard error across queries. (b) Per-query hard-ranking margins before and after correction. The diagonal denotes no change, and the green region denotes improvement.

Foundation references. Table 5 compares which frozen affinities enter the subtractive score while keeping the task model and modality-weight fitting unchanged. With learned task-modality weights but no reference term, mAP is 78.2%. Adding only the global or local reference yields 83.5% and

81.1%, respectively. Combining both achieves the highest mAP of 83.7%, with 86.7% Rank-1, 93.5% Rank-5, and 95.7% Rank-10. Compared with no reference term, mAP improves by 5.5 percentage points, showing that global and local affinities together improve retrieval.

Table 5: Ablation of foundation-affinity references on RGBNT201 (%).

Reference support	mAP	Rank-1	Rank-5	Rank-10
$\varnothing$	78.2	79.9	87.1	91.4
G	83.5	**86.7**	92.3	95.0
L	81.1	83.7	91.4	94.3
$G + L$	**83.7**	**86.7**	**93.5**	**95.7**

Parameter sensitivity. Figure 3 varies α and δ on RGBNT201 while fixing the other settings. Across the plotted α settings, mAP ranges from 82.6% to 83.7%, and $\alpha = 1$ yields the highest mAP and Rank-1. Across the margin settings, mAP ranges from 82.7% to 83.7%, peaking at $\delta = 0.2$.

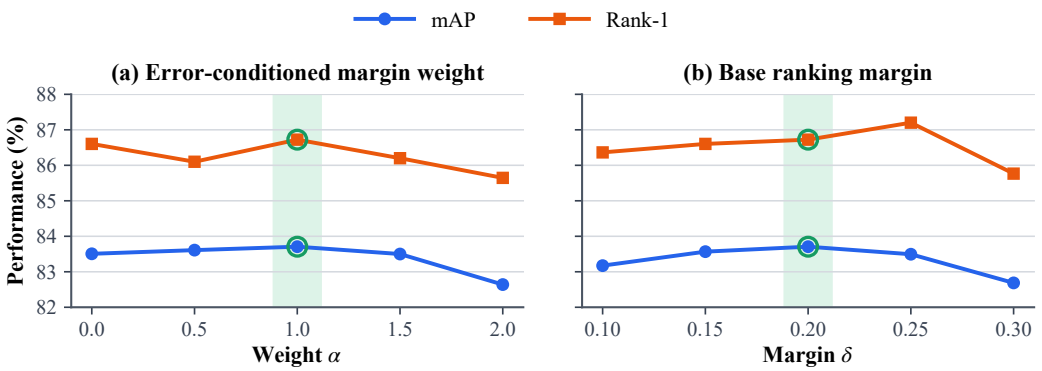

Figure 3: Sensitivity to the error-conditioned margin coefficient α and ordinary margin δ on RGBNT201. Green regions and hollow markers indicate the default values.

4.5 Efficiency Analysis

On RGBNT201, score correction increases mAP from 78.2% to 83.7% and Rank-1 from 79.9% to 86.7%, while SAFR and its uncorrected task encoder both have 86.50M trainable parameters. Including the frozen reference forward pass raises feature-extraction cost from 35.13 to 64.45 GFLOPs per 256×128 RGB-NIR-TIR triplet. On an NVIDIA GeForce RTX 4090 D with BF16, batch-one latency increases from 5.34 to 11.36 ms; pairwise score correction adds 0.0115M multiply-adds per query-gallery pair. Full comparisons appear in Appendix A.

5 Conclusion

We presented SAFR for incorporating frozen foundation affinities into multi-modal identity retrieval. Foundation-error-aware ranking complements standard ReID supervision by strengthening task margins when the frozen reference reverses a positive-negative ordering. Sparse affinity correction then estimates and selects global-local reference contributions on training identities to refine the fixed retrieval score. Experiments on three person and vehicle benchmarks, together with component and mechanism analyses, support the complementary roles of representation learning and score correction. The frozen encoder adds feature-extraction cost; reducing this overhead while retaining useful reference relations is a direction for future work.

AI USE STATEMENT

Generative AI tools assisted with manuscript writing and polishing by suggesting clearer phrasing, improving grammar, and checking terminology consistency. They were also used to search for and identify related studies on multi-modal ReID, visual foundation models, and retrieval-score refinement. The authors checked the suggested references against the original publications, reviewed the AI-assisted text, and verified reported numerical results against saved experiment outputs or author-provided remeasurements. The authors take responsibility for the final submission.

REFERENCES

Song Bai, Peng Tang, Philip H. S. Torr, and Longin Jan Latecki. Re-ranking via metric fusion for object retrieval and person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 740–749, 2019.

Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. Deep metric learning to rank. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1861–1870, 2019.

Jennifer Crawford, Haoli Yin, Luke McDermott, and Daniel Cummings. UniCat: Crafting a stronger fusion baseline for multimodal re-identification. In *NeurIPS Workshop on Unifying Representations in Neural Models*, 2023.

Wenjiao Dong, Xi Yang, De Cheng, Nannan Wang, and Xinbo Gao. Escaping modal interactions: An efficient DESANet for multi-modal object re-identification. *IEEE Transactions on Image Processing*, 34:5068–5083, 2025. doi: 10.1109/TIP.2025.3592575.

Yingying Feng, Jie Li, Jie Hu, Yukang Zhang, Lei Tan, and Jiayi Ji. MDReID: Modality-decoupled learning for any-to-any multi-modal object re-identification. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Yingying Feng, Jie Li, Chi Xie, Lei Tan, and Jiayi Ji. Multi-modal object re-identification via sparse mixture-of-experts. In *Proceedings of the International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16796–16807, 2025b.

Ya Gao, Shihao Li, Zhaojun Liu, Aihua Zheng, Chenglong Li, and Jin Tang. Chain-of-thought guided multi-modal object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37705–37714, 2026.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15013–15022, 2021.

Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

He Li, Mang Ye, Ming Zhang, and Bo Du. All in one framework for multimodal re-identification in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17459–17469, 2024.

Hongchao Li, Chenglong Li, Xianpeng Zhu, Aihua Zheng, and Bin Luo. Multi-spectral vehicle re-identification: A challenge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11345–11353, 2020. doi: 10.1609/aaai.v34i07.6796.

Hongchao Li, Guangxing Liu, Xixi Wang, Baihe Liang, and Yonglong Luo. PEFT-BoA: Parameter-efficient fine-tuning with bag-of-adapters for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6127–6135, 2026a. doi: 10.1609/aaai.v40i8.37537.

Shihao Li, Huaibo Huang, Junxian Duan, Aihua Zheng, Jin Tang, and Jixin Ma. NEXT: Multi-grained mixture of experts via text-modulation for multi-modal object re-identification. *IEEE Transactions on Information Forensics and Security*, 2026b. doi: 10.1109/TIFS.2026.3729508.

Shihao Li, Chenglong Li, Aihua Zheng, Jin Tang, and Bin Luo. ICPL-ReID: Identity-conditional prompt learning for multi-spectral object re-identification. *IEEE Transactions on Multimedia*, 28: 599–613, 2026c. doi: 10.1109/TMM.2025.3623556.

Siyuan Li, Li Sun, and Qingli Li. CLIP-ReID: Exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1405–1413, 2023. doi: 10.1609/aaai.v37i1.25225.

Minghui Lin, Shu Wang, Xiang Wang, Jianhua Tang, Longbin Fu, Zhengrong Zuo, and Nong Sang. DMPT: Decoupled modality-aware prompt tuning for multi-modal object re-identification. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 2103–2112, 2025.

Yangyang Liu, Yuhao Wang, and Pingping Zhang. Signal: Selective interaction and global-local alignment for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7359–7367, 2026. doi: 10.1609/aaai.v40i9.37674.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019.

Xuanhao Qi, Tom H. Luan, Yukang Zhang, Jinkai Zheng, Zhou Su, Shuwei Li, and Lei Tan. FUSE: Frequency-domain unification and spectral energy alignment for multi-modal object re-identification. In *Proceedings of the International Conference on Machine Learning*, 2026.

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 *Proceedings of the International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Xixi Wan, Aihua Zheng, Bo Jiang, Beibei Wang, Chenglong Li, and Jin Tang. UGG-ReID: Uncertainty-guided graph model for multi-modal object re-identification. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yuhao Wang, Xuehu Liu, Pingping Zhang, Hu Lu, Zhengzheng Tu, and Huchuan Lu. TOP-ReID: Multi-spectral object re-identification with token permutation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5758–5766, 2024a. doi: 10.1609/aaai.v38i6.28388.

Yuhao Wang, Xuehu Liu, Tianyu Yan, Yang Liu, Aihua Zheng, Pingping Zhang, and Huchuan Lu. MambaPro: Multi-modal object re-identification with mamba aggregation and synergistic prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8150–8158, 2025a. doi: 10.1609/aaai.v39i8.32879.

Yuhao Wang, Yang Liu, Aihua Zheng, and Pingping Zhang. DeMo: Decoupled feature-based mixture of experts for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8141–8149, 2025b. doi: 10.1609/aaai.v39i8.32878.

Yuhao Wang, Yongfeng Lv, Pingping Zhang, and Huchuan Lu. IDEA: Inverted text with cooperative deformable aggregation for multi-modal object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29701–29710, 2025c.

Zi Wang, Chenglong Li, Aihua Zheng, Ran He, and Jin Tang. Interact, embed, and enlarge: Boosting modality-specific representations for multi-modal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2633–2641, 2022. doi: 10.1609/aaai.v36i3.20165.

Zi Wang, Huaibo Huang, Aihua Zheng, and Ran He. Heterogeneous test-time training for multi-modal person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5850–5858, 2024b. doi: 10.1609/aaai.v38i6.28398.

Di Wu, Zhihui Liu, Zihan Chen, Shenglong Gan, Kaiwen Tan, Qin Wan, and Yaonan Wang. LRMM: Low rank multi-scale multi-modal fusion for person re-identification based on RGB-NI-TI. *Expert Systems with Applications*, 263:125716, 2025. doi: 10.1016/j.eswa.2024.125716.

Xingguo Xu, Zhanyu Liu, Weixiang Zhou, Yuansheng Gao, Junjie Cao, Yuhao Wang, Jixiang Luo, and Dell Zhang. STMI: Segmentation-guided token modulation with cross-modal hypergraph interaction for multi-modal object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11433–11441, 2026. doi: 10.1609/aaai.v40i14.38125.

Haoli Yin, Jiayao Li, Eva Schiller, Luke McDermott, and Daniel Cummings. GraFT: Gradual fusion transformer for multimodal re-identification. *arXiv preprint arXiv:2310.16856*, 2023.

Zhi Yu, Zhiyong Huang, Mingyang Hou, Jiaming Pei, Yan Yan, Yushi Liu, and Daming Sun. Representation selective coupling via token sparsification for multi-spectral object re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(4):3633–3648, 2025. doi: 10.1109/TCSVT.2024.3509817.

Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17117–17126, 2024.

Shizhou Zhang, Wenlong Luo, De Cheng, Yinghui Xing, Guoqiang Liang, Peng Wang, and Yanning Zhang. Prompt-based modality alignment for effective multi-modal object re-identification. *IEEE Transactions on Image Processing*, 34:2450–2462, 2025. doi: 10.1109/TIP.2025.3556531.

Aihua Zheng, Zi Wang, Zihan Chen, Chenglong Li, and Jin Tang. Robust multi-modality person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 3529–3537, 2021. doi: 10.1609/aaai.v35i4.16467.

Aihua Zheng, Xianpeng Zhu, Zhiqi Ma, Chenglong Li, Jin Tang, and Jixin Ma. Cross-directional consistency network with adaptive layer normalization for multi-spectral vehicle re-identification and a high-quality benchmark. *Information Fusion*, 100:101901, 2023.

Aihua Zheng, Zhiqi Ma, Yongqi Sun, Zi Wang, Chenglong Li, and Jin Tang. Flare-aware cross-modal enhancement network for multi-spectral vehicle re-identification. *Information Fusion*, 116:102800, 2025. doi: 10.1016/j.inffus.2024.102800.

Aihua Zheng, Pengyu Li, Zi Wang, and Jin Tang. Progressive multi-modal knowledge distillation for multi-spectral object re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13351–13359, 2026a. doi: 10.1609/aaai.v40i16.38338.

Aihua Zheng, Zhaojun Liu, Xixi Wan, Chenglong Li, Jin Tang, and Yan Yan. ProxyTTT: Proxy-driven test-time training for multi-modal re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13342–13350, 2026b. doi: 10.1609/aaai.v40i16.38337.

Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1318–1327, 2017.

SUPPLEMENTARY MATERIALS

Additional implementation details. The task encoder applies a separate residual adapter to the shared DINOv3 output for each of RGB, NIR, and TIR, allowing modality-dependent feature adjustment without duplicating the backbone. Each adapter uses LayerNorm, a 768-to-64-to-768 bottleneck with GELU, and a residual connection; the adapted feature is then ℓ_2-normalized. Label smoothing is 0.1, the triplet margin is 0.3, and $\epsilon = 10^{-12}$. For task-weight regression, each eligible training anchor contributes up to four same-identity positives from other camera or scene groups and up to four sampled different-identity negatives per selected positive. Reference coefficients use all different-identity training pairs.

A EFFICIENCY ANALYSIS

Parameters and computation. Table 6 gives the full model-complexity and RGBNT201-accuracy comparison. SAFR exceeds DeMo by 4.7 mAP and 4.4 Rank-1 points, and IDEA by 3.5 and 4.6 points, respectively.

Table 6: Model complexity and RGBNT201 performance. Lower values are better for parameters and GFLOPs. DeMo complexity follows its original paper. Our task encoder and SAFR are profiled with fvcore on a 256×128 RGB-NIR-TIR triplet; figures for external methods may follow different counting protocols. SAFR FLOPs include the frozen reference forward pass, whereas its frozen parameters are excluded from the trainable count.

Method	Params. (M)	GFLOPs	mAP	Rank-1
TOP-ReID (Wang et al., 2024a)	324.52	35.52	72.3	76.6
DeMo (Wang et al., 2025b)	98.79	**35.10**	79.0	<u>82.3</u>
IDEA (Wang et al., 2025c)	<u>91.67</u>	43.73	<u>80.2</u>	82.1
Task encoder (uncorrected)	**86.50**	<u>35.13</u>	78.2	79.9
SAFR (ours)	**86.50**	64.45	**83.7**	**86.7**

B REFERENCE-LAYER ANALYSIS

Table 7 compares five global-local extraction-block pairs with the same task encoder and training-only fitting protocol. The 8/10 pair reaches the highest mAP at 83.7%, with 86.7% Rank-1; the 12/12 pair obtains 82.7% mAP and 85.5% Rank-1.

Table 7: Effect of reference-feature extraction blocks on RGBNT201 (%).

Global block	Local block	mAP	Rank-1	Rank-5	Rank-10
8	8	83.4	86.4	92.8	95.2
8	10	**83.7**	**86.7**	**93.5**	**95.7**
10	10	82.5	**86.7**	91.5	94.3
10	12	82.5	**86.7**	91.5	94.3
12	12	82.7	85.5	91.4	93.9

C ESTIMATION OF CALIBRATION COEFFICIENTS

We detail how the calibration coefficients in Eq. (13) are estimated for a fixed reference subset $\mathcal{S} \subseteq \{G, L\}$. The task encoder, reference features, and modality weights remain fixed. The problem therefore involves at most two scalar coefficients, rather than further neural-network training.

Fitting centered similarities. For each training anchor i, we use its different-identity candidates and center both the task scores and each reference column:

$$\widetilde{\mathbf{k}}_i = H_i \mathbf{k}^i_{\text{task}}, \qquad \widetilde{X}_i = H_i X_i^{\mathcal{S}}.$$

Subtracting these per-anchor means removes query-specific offsets. The coefficients are then fitted to the remaining variation in task scores:

$$\boldsymbol{\beta}^*_{\mathcal{S}} = \arg \min_{0 \le \boldsymbol{\beta} \le \beta_{\max} \mathbf{1}} \sum_i \|\widetilde{\mathbf{k}}_i - \widetilde{X}_i \boldsymbol{\beta}\|_2^2, \qquad \beta_{\max} = 3.$$

This is the same objective as Eq. (13). Each different-identity pair contributes to the sum; no additional per-anchor averaging is applied.

Solving the bounded problem. Writing $d = |\mathcal{S}|$, define

$$Q = \sum_i \widetilde{X}_i^\top \widetilde{X}_i \in \mathbb{R}^{d \times d}, \qquad q = \sum_i \widetilde{X}_i^\top \widetilde{\mathbf{k}}_i \in \mathbb{R}^d.$$

After dropping a constant independent of $\boldsymbol{\beta}$, the objective becomes $\boldsymbol{\beta}^\top Q \boldsymbol{\beta} - 2q^\top \boldsymbol{\beta}$. Since Q is positive semidefinite, the problem is convex. Its feasible region is an interval for one reference and a rectangle for two references.

We solve this small problem by checking the interior and its boundaries. An interior candidate satisfies $Q\boldsymbol{\beta} = q$. Boundary candidates are obtained by fixing one or more coefficients at 0 or $\beta_{\max}$ and solving for the remaining coefficients. We discard infeasible candidates and retain the one with the lowest objective value. Each coefficient can be free, zero, or at its upper bound, so at most nine cases are considered. Comparing the feasible interior, edge, and vertex candidates yields a global minimum; the minimizing coefficients need not be unique. The implementation accumulates centered statistics in double precision, uses least-squares linear solves for singular systems, and allows a small numerical tolerance when checking the bounds.

Using the fitted coefficients. The resulting coefficients define the subtractive score correction in Eq. (14). Coefficient fitting is performed on the fitting identities within each cross-validation fold. After the reference subset is selected using held-out training identities, the coefficients are refitted on all training identities and fixed for retrieval. If $\mathcal{S} = \varnothing$, no coefficients are needed and the score remains K_{task}. Thus, the solver estimates the strength of a given reference subset; the separate selection procedure determines which references to retain.

D Qualitative Analysis

Retrieval examples. Figure 4 compares plain ReID and full SAFR for two RGBNT201 queries. The baseline retrieves zero and one correct match in their respective top-10 lists, whereas full SAFR retrieves ten for each. Their APs rise from 38.0% to 97.2% and from 25.1% to 98.0%, respectively. The examples show how the complete method moves true matches ahead of visually similar distractors. Figure 5 makes the same comparison for two MSVR310 vehicle queries. The number of correct top-5 matches rises from two to five for the first query and from three to five for the second. Their APs rise from 42.1% to 67.0% and from 55.3% to 89.5%, respectively. These selected person and vehicle examples illustrate the complete method's effect on individual rankings; the aggregate comparisons in Section 4.4 measure its effect across all queries. Both figures use the models and training-fitted correction associated with the reported main results.



Figure 4: Two-stage multi-modal person rank lists on RGBNT201. Two queries are shown side by side for (a) plain ReID and (b) full SAFR. In each RGB, NIR, and TIR row, the leftmost image is the query; numbers 1–10 mark retrieval ranks. Green and red borders mark correct and incorrect matches.

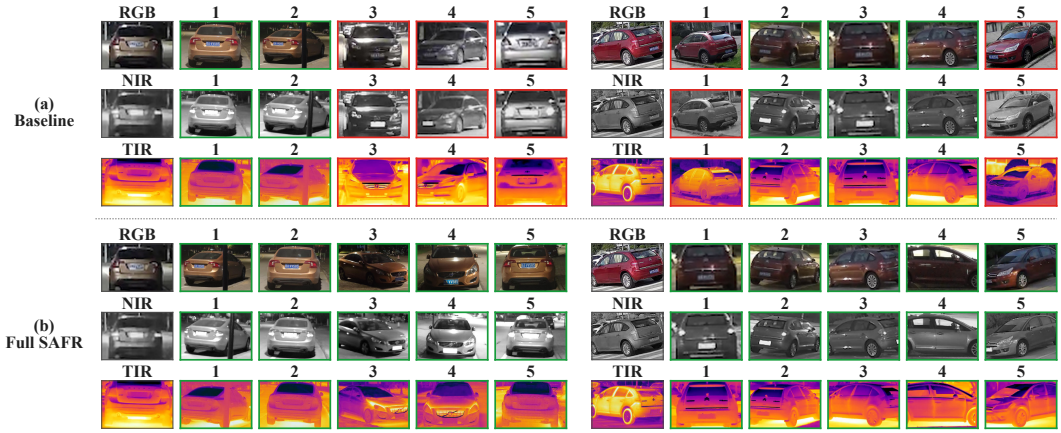

Figure 5: Two-stage multi-modal vehicle rank lists on MSVR310. Two queries are shown side by side for (a) plain ReID and (b) full SAFR. In each RGB, NIR, and TIR row, the leftmost image is the query; numbers 1–5 mark retrieval ranks. Green and red borders mark correct and incorrect matches.