TRIDA: TRI-MODAL RESIDUAL-AWARE INTER-MODAL DIVERGENCE ARBITRATION FOR MULTI-SPECTRAL OBJECT RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-spectral object re-identification (ReID) aims to match identities across heterogeneous spectral observations. This matching capability is critical under varying illumination and adverse sensing conditions. Cross-spectral interaction is essential for exploiting complementary spectral cues, yet it is not cost-free: token selection discards potentially discriminative cues at unselected positions, all-token retention propagates modality-specific interference, and unconstrained interaction further incurs representation drift and gradient conflicts. We propose Tri-modal Residual-aware Inter-modal Divergence Arbitration (TRIDA), which addresses these challenges through three coordinated designs. Specifically, GRACE retains all tokens to preserve identity cues and reorganizes cross-modal information via channel-group cyclic permutation, while cross-modal CLS differences adaptively modulate CLS attention and residual scaling to mitigate modality-specific interference. GRASP then counters representation drift by regularizing group-wise deviations toward a sample-wise cross-modal center with progressive scheduling, while preserving discriminative modality-specific information. Finally, MCGA resolves conflicting gradients: it probes modality-wise gradients from the identity-discriminative and image-text objectives on the trainable GRACE enhancement parameters, and applies conflict-severity-adaptive, order-independent arbitration, leaving all unprobed gradient contributions intact. Through empirical evaluations on five multi-spectral object ReID benchmarks, TRIDA demonstrates superior object retrieval performance while requiring only modest parameter overhead.

1 INTRODUCTION

Object re-identification (ReID) matches the same identity across non-overlapping cameras. Despite progress in RGB-based ReID, visible-spectrum imagery degrades under weak illumination, glare, or occlusion (Xi, 2025; Zhang et al., 2024). Multi-spectral sensing addresses this: RGB preserves color and texture, near-infrared (NIR) provides reflectance cues under low light, and thermal infrared (TIR) captures radiation independent of ambient light. However, these spectra differ markedly in appearance statistics, resolution, and identity-discriminative cues. Given these complementary yet heterogeneous spectral characteristics, the central challenge is not simply to fuse multiple spectra, but to establish effective cross-spectral interaction while preserving modality-specific identity cues and maintaining stable joint optimization.

As illustrated in Fig. 1, cross-spectral modeling in multi-spectral ReID has evolved along three representative paradigms, each exposing a distinct limitation that TRIDA is proposed to address. Early fusion and alignment methods (Fig. 1(a)) encode each spectrum with a dedicated branch and then concatenate or group the characteristics of the modality into a unified embedding space (Zheng et al., 2023b; 2021). This paradigm is simple, but treats shared and modality-unique cues indiscriminately: merging

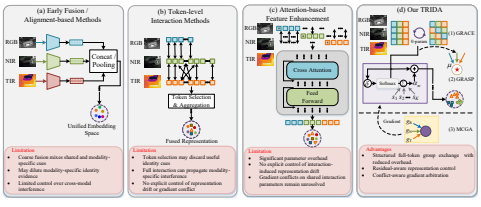

Figure 1: Multi-spectral ReID paradigms.

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.

heterogeneous features at the embedding level risks feature contamination, and aggressive pooling may erase discriminative modality-specific evidence rather than preserving it.

Token-level interaction methods (Fig. 1(b)) are predominantly Transformer-based. They move the interaction to a finer granularity. TOP-ReID (Wang et al., 2024a) permutes tokens to reconstruct complementary features, while Magic Tokens (Zhang et al., 2024), RSCNet (Yu et al., 2024), PromptMA (Zhang et al., 2025b), and IDEA (Wang et al., 2025c) introduce selective interaction through spatial-frequency selection, sparse coupling, prompt exchange, or deformable aggregation. Although such a local interaction is more expressive, it introduces two coupled problems. First, token selection keeps only a subset of positions and can discard potentially discriminative cues, while the retained tokens may still transmit clutter, artifacts, and redundancy across modalities.

To strengthen the interacted descriptors, a third paradigm adopts attention-based feature enhancement (Fig. 1(c)), stacking cross-attention and feed-forward blocks that refine each modality through cross-modal modulation (Wang et al., 2022; Feng et al., 2023; Jiang et al., 2022). While improving discriminability, these blocks introduce additional trainable transformations and computational overhead and still provide no explicit control over interaction-induced representation drift, *i.e.*, how far interacted features move away from a stable identity structure.

A less visible problem is the optimization conflict, which arises when the modality-wise objectives share the interaction parameters. OGM-GE (Peng et al., 2022) shows that joint multimodal training can under-optimize weaker modalities, DGL (Wei et al., 2025) identifies interference between modality encoders and the fusion module, and PCGrad (Yu et al., 2020) projects conflicting gradients; yet cross-spectral ReID typically leaves such conflicts implicit for the shared interaction parameters. Therefore, we propose TRIDA (Fig. 1(d)), whose three modules directly address the limitations above. GRACE replaces token selection with a parameter-free cyclic group exchange: it partitions channels into disjoint groups, keeps designated groups native to each modality, and retains *all* tokens, so that complementary cues propagate without discarding positions or erasing modality-specific evidence. GRASP then regularizes each exchanged group to a sample-wise cross-modal center with progressive scheduling, explicitly suppressing the fusion drift that token-interaction methods leave uncontrolled. Finally, MCGA probes the modality-wise gradients of the trainable GRACE enhancement parameters and applies conflict-severity-adaptive arbitration, coordinating competing optimization directions while leaving all unprobed gradient contributions intact. These modules jointly form TRIDA, and our main contributions are summarized as follows:

- We develop GRACE, a structured cross-modal exchange module that partitions channels into disjoint groups and applies a deterministic, modality-cyclic permutation. The exchange is parameter-free, retains *all* tokens without token selection, and keeps designated groups native to preserve modality-specific cues while enabling controlled complementary propagation.
- We introduce GRASP, a group-wise residual anchoring loss with linear warm-up that constrains each exchanged group's deviation from a sample-wise cross-modal center. Using the same grouping as GRACE, it suppresses interaction-induced drift without enforcing exact cross-modal equality or sacrificing discriminative modality-specific representations.
- We propose MCGA, a conflict-triggered gradient arbitration mechanism that probes per-modality gradients on the trainable GRACE enhancement parameters, detects conflicts by cosine thresholding, and resolves them with order-independent symmetric projection and severity-adaptive blending, while retaining all unprobed gradient contributions.

2 Related Work

Multi-spectral object re-identification. Multi-spectral ReID differs from visible-infrared cross-modality matching: synchronized RGB, NIR, and TIR observations are available on both the query and gallery sides, requiring a method to simultaneously bridge the spectral gap and aggregate complementary evidence. Existing methods can be distinguished by where the cross-spectral interaction occurs. One line performs interaction at the embedding level, fusing dedicated per-spectrum encoders through progressive or consistency-constrained fusion (Zheng et al., 2023b; 2021; 2023a; Crawford et al., 2023). With Vision Transformers (Dosovitskiy et al., 2021), interaction shifts to tokens or sequences, including cyclic token permutation (Wang et al., 2024a), object-centric token

selection (Zhang et al., 2024), sparse or gradual coupling (Yu et al., 2024; Yin et al., 2023; Liu et al., 2026; Xu et al., 2026), and efficient sequence modeling (Wang et al., 2025a); cross-attention and feed-forward stacks instead refine each spectrum through heavy modality-mixing blocks (Wang et al., 2022; Feng et al., 2023; Jiang et al., 2022). A further set of mechanisms extends the setting itself through prompt exchange and semantic inversion (Zhang et al., 2025b; Wang et al., 2025c), any-to-any or missing-modality inference (Feng et al., 2025a; Xi, 2025), and mixture-of-experts routing (Wang et al., 2025b; Feng et al., 2025b), with semantics-guided variants pursuing cross-modal consistency at the modulation level (Gao et al., 2026; Li et al., 2025a; Zhang et al., 2025a). These designs determine which forward signals are exchanged, but they neither bound the deviation of interacted representations from an identity-consistent structure nor arbitrate the modality-wise gradients that update shared interaction parameters.

Vision-language adaptation for ReID. CLIP (Radford et al., 2021) provides transferable visual-semantic concepts from image-text supervision, which prompt learning (Zhou et al., 2022) and lightweight adapters (Chen et al., 2022) can specialize with few tuned parameters. Building on this paradigm, CLIP-ReID (Li et al., 2023) learns identity-specific prompts without manual text annotations, while ICPL-ReID (Li et al., 2025b) extends this idea to multi-spectral ReID with online prompts, identity prototypes, and modality adapters; factorized prompts and multi-scope adapters further decouple spectral style from object content (Lin et al., 2025; Li et al., 2026a;b). Such semantic alignment determines the target semantics that interacted features should approach, but it does not specify which feature groups should be exchanged, how strongly exchanged features should be anchored, or how shared parameters should respond to incompatible modality gradients.

Gradient conflicts in shared-parameter learning. When multiple objectives share parameters, conflicting gradients with negative inner products make a descent step for one objective degrade another. Gradient-based multi-task learning mitigates this via projection (Yu et al., 2020), conflict-averse direction search (Liu et al., 2021), bargaining-based aggregation (Navon et al., 2022), or loss reweighting (Chen et al., 2018), whereas multimodal-training methods modulate modality gradients and interference between the encoder and fusion module (Peng et al., 2022; Wei et al., 2025). These approaches operate on whole-task gradients over the full parameter space, and sequential projection is further sensitive to task order. Our setting instead updates a small shared interaction subspace alongside a largely frozen backbone and prompt-prototype terms. MCGA probes only the modality-discriminative and image-text gradients on the trainable GRACE parameters, preserves all unprobed gradient contributions, and computes pairwise corrections from the original gradients using symmetric, severity-adaptive blending. Consequently, arbitration is confined to the shared subspace, is order-independent, and recovers the ordinary update when no conflict is detected.

3 METHODOLOGY

We present Tri-modal Residual-aware Inter-modal Divergence Arbitration (TRIDA), built on the prompt-prototype learning framework of ICPL-ReID. As illustrated in Fig. 2, TRIDA comprises three cooperative components: (1) Group-wise Residual Adaptive Cross-modal Exchange (GRACE) for structured inter-modal information transfer, (2) Group-wise Residual Anchoring with Soft Progression (GRASP) for identity-preserving regularization, and (3) Modality-Consensus Gradient Arbitration (MCGA) for conflict-aware optimization.

3.1 GROUP-WISE RESIDUAL ADAPTIVE CROSS-MODAL EXCHANGE

For each training sample i, we are given synchronized images $\{I_m^i\}_{m=0}^{M-1}$ and an identity label y_i, where M denotes the number of modalities ($M = 3$ for RGB, near-infrared, and thermal infrared). Following ICPL-ReID, we use a CLIP ViT-B/16 image encoder with lightweight modality adapters and retain its prompt–prototype learning framework. The encoder produces a token sequence $\mathbf{Z}_m^i \in \mathbb{R}^{(L+1)\times D}$, consisting of one CLS token and L patch tokens. As illustrated in Fig. 3, GRACE combines a parameter-free redistribution of channel groups with a learnable residual transformation. The exchange determines which

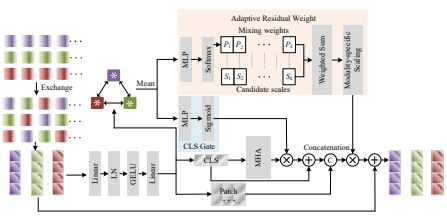

Figure 3: Details of the GRACE module.

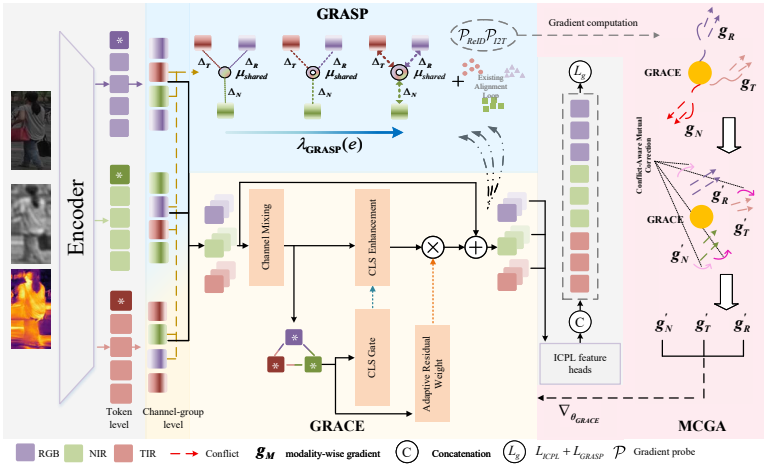

Figure 2: TRIDA: GRACE (cyclic group exchange and bounded enhancement), GRASP (cross-modal anchoring), and MCGA (backward gradient arbitration); only GRACE at inference.

modality supplies each feature group; the subsequent transformation adapts these recombined features and introduces cross-modal interaction between CLS tokens. A cross-modal difference-guided dual-level modulation mechanism regulates both the CLS-attention injection and residual-branch scaling based on sample-specific modality discrepancies.

Group-wise exchange. Let $\mathbf{Z}_m$ denote the feature of modality m before exchange. We partition the channel dimension into G contiguous groups of size $d = D/G$ (with $G \mid D$). For modality m and group g, the exchanged feature is:

$$Z^{\text{ex}}_{m,g} = Z_{(m+g) \bmod M, g}, \quad g = 0, \ldots, G-1 \tag{1}$$

Concatenating groups along the channel dimension yields $\mathbf{Z}^{ex}_m \in \mathbb{R}^{(L+1)\times D}$. All groups are read from original modality features, so the exchange is simultaneous rather than sequential. Groups satisfying $g \bmod M = 0$ remain in their original modality; for example, with $M = 3$ and $G = 4$, groups 0 and 3 are retained while groups 1 and 2 follow cyclic shifts. This operation is parameter-free and redistributes existing information across modality streams, leaving adaptation to the learnable enhancement stage.

Discrepancy-conditioned enhancement. A shared token-wise channel mapping produces $\mathbf{H}_m = \phi(\mathbf{Z}^{\text{ex}}_m)$, where ϕ consists of a linear layer, LayerNorm, GELU, and a second linear layer. Let $\mathbf{h}_m = \mathbf{H}_m[0, :]$ denote its CLS representation. We summarize cross-modal differences using the mean pairwise absolute difference:

$$\boldsymbol{\delta} = \frac{2}{M(M-1)} \sum_{0 \le m < n < M} |\mathbf{h}_m - \mathbf{h}_n|, \quad M \ge 2 \tag{2}$$

The vector $\boldsymbol{\delta} \in \mathbb{R}^D$ conditions both the strength of CLS interaction and residual scaling, describing current feature differences without explicitly assigning them to nuisance variation or useful modality-specific information. We stack CLS representations into $\mathbf{H}_{\text{cls}} \in \mathbb{R}^{M\times D}$ and apply Multi-Head Self-Attention across modalities:

$$\mathbf{V} = \sigma\left(\psi_{\text{gate}}(\boldsymbol{\delta})\right) \text{MHA}\left(\mathbf{H}_{\text{cls}}, \mathbf{H}_{\text{cls}}, \mathbf{H}_{\text{cls}}\right) \tag{3}$$

where ψ_{gate} is a two-layer perceptron and $\sigma(\cdot)$ denotes the sigmoid function, producing one scalar gate per sample shared across modalities and channels. Cross-modal attention is applied only to CLS tokens rather than full patch-token sequences.

To obtain a sample-adaptive residual scale, the discrepancy $\boldsymbol{\delta}$ is passed through a two-layer perceptron (ReLU) and softmax, producing assignment weights p_k over K preset reference scales s_k, with $\sum_{k=1}^{K} p_k = 1$ and $p_k \ge 0$. A base residual scale is thus obtained as their convex combination. We

further introduce a modality-specific learnable scalar a_m for amplitude adjustment:

$$\alpha_m = \mathrm{clip}\left(\mathrm{clip}\left(a_m, a_{\min}, a_{\max}\right) \sum_{k=1}^{K} p_k s_k, \alpha_{\min}, \alpha_{\max}\right) \tag{4}$$

This formulation yields a continuous, sample-adaptive residual strength rather than a single discrete candidate: the set s only provides reference scales, the final strength is jointly determined by cross-modal discrepancy and modality-specific adjustment, and K can be varied to control the scale resolution. The clipping operations bound the coefficient α_m, not the residual-feature norm. Let $\tilde{\mathbf{V}}_m \in \mathbb{R}^{(L+1)\times D}$ contain the m-th row of $\mathbf{V}$ at the CLS position and zeros at all patch positions. The enhanced sequence is:

$$\mathbf{Z}_m^{en} = \mathbf{Z}_m^{\mathrm{ex}} + \alpha_m \left(\mathbf{H}_m + \tilde{\mathbf{V}}_m\right) \tag{5}$$

where $\mathbf{Z}_m^{en}$ denotes the enhanced representation of modality m. The channel mapping acts on each token independently, whereas attention contributes only to the CLS update. The gate in Eq. (3) scales the attention contribution, while α_m scales the entire residual branch; the exchanged representation remains as the direct residual path.

3.2 Group-wise Residual Anchoring with Soft Progression

Feature exchange changes the composition of each modality stream but does not explicitly constrain differences between resulting representations. GRASP penalizes group-wise deviations of exchanged CLS features from their cross-modal mean. It is applied before enhancement, so the regularizer acts on exchanged representations rather than attention-adjusted outputs, as illustrated in Fig. 4.

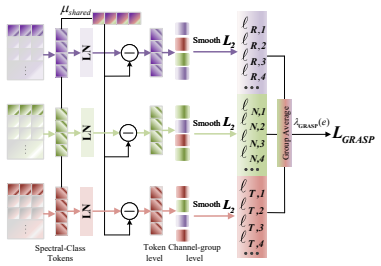

Figure 4: GRASP: group-wise deviation penalty on exchanged CLS features.

For sample i, we first normalize each exchanged CLS feature across its full channel dimension:

$$\begin{aligned} \mathbf{x}_m^i &= \mathrm{LN}_D\left(\mathbf{Z}_m^{\mathrm{ex},i}[0,:]\right), \\ \boldsymbol{\mu}^i &= \frac{1}{M}\sum_{m=0}^{M-1} \mathbf{x}_m^i, \\ \mathbf{r}_m^i &= \mathbf{x}_m^i - \boldsymbol{\mu}^i \end{aligned} \tag{6}$$

where LN_D denotes layer normalization without learnable affine parameters, and the mean $\boldsymbol{\mu}^i$ remains differentiable (not detached from the computation graph). Partitioning $\mathbf{r}_m^i$ into the same G channel groups yields:

$$\mathcal{L}_{\mathrm{GRASP}} = \frac{1}{BMG}\sum_{i=1}^{B}\sum_{m=0}^{M-1}\sum_{g=0}^{G-1}\sqrt{\left\|\mathbf{r}_{m,g}^{(i)}\right\|_2^2 + \epsilon} \tag{7}$$

The smoothed group norm grows approximately linearly with deviation, imposing soft cross-modal consistency rather than exact equality. We ramp its weight during the first E_w epochs:

$$\lambda_G(e) = \lambda_G^{\max}\min\left(1, \frac{e}{E_w}\right). \tag{8}$$

3.3 Modality-Consensus Gradient Arbitration

Modality-specific objectives share the enhancement parameters but may produce conflicting gradients: when $\langle \mathbf{g}_m, \mathbf{g}_n \rangle < 0$, a step along $-\mathbf{g}_m$ increases the other objective to first order. MCGA resolves such pairwise conflicts during gradient aggregation without altering the scalar training objective.

Modality-specific gradients. Let θ_f denote the trainable GRACE enhancement parameters selected for arbitration. From the same forward pass, we compute:

$$\mathbf{g}_m = \nabla_{\theta_f} \left[\omega_m \left(\mathcal{L}^m_{\text{ReID}} + \lambda_{\text{i2t}} \mathcal{L}^m_{\text{i2t}} \right) \right] \tag{9}$$

where ω_m is the modality weight and λ_{i2t} balances the ReID and image-text alignment losses. Gradients are flattened in a consistent parameter order. The image-to-prototype objective is excluded from this auxiliary gradient computation because its backward implementation updates the prototype memory; it remains part of the ordinary training objective. The prototype-text terms use detached prototype features and have no gradient path to θ_f.

Conflict-dependent correction. Define $q_m = \max\left(\|\mathbf{g}_m\|_2^2, \epsilon_g\right)$, where $\epsilon_g > 0$ is a numerical stability constant. Pairwise agreement c_{mn} is measured by:

$$c_{mn} = \frac{\langle \mathbf{g}_m, \mathbf{g}_n \rangle}{\sqrt{q_m q_n}} \tag{10}$$

where c_{mn} measures the cosine agreement between modalities m and n. We regard a pair as conflicting when $c_{mn} < \tau_c$, where $\tau_c < 0$ leaves weak negative agreement untreated. For each modality, let $\mathcal{C}_m = \{n \neq m : c_{mn} < \tau_c\}$. The corrected gradient is:

$$\hat{\mathbf{g}}_m = \mathbf{g}_m - \rho_p \sum_{n \in \mathcal{C}_m} \frac{\langle \mathbf{g}_m, \mathbf{g}_n \rangle}{q_n} \mathbf{g}_n \tag{11}$$

where ρ_p controls the correction strength. All corrections are computed from the original gradients and aggregated simultaneously, making the result order-independent without requiring nonnegative pairwise inner products. Let $c_{\min} = \min_{m<n} c_{mn}$, the most negative pairwise similarity, determine the mixing strength:

$$\beta = \beta_0 \text{clip}\left(\frac{\tau_c - c_{\min}}{\max(\epsilon_g, 1 + \tau_c)}, 0, 1 \right) \tag{12}$$

where β_0 is the maximum blending coefficient. With $\mathbf{g}_P = \sum_m \mathbf{g}_m$, the mixed gradient is:

$$\mathbf{g}_A = (1 - \beta)\mathbf{g}_P + \beta \sum \hat{\mathbf{g}}_m \tag{13}$$

where $\mathbf{g}_A$ is the conflict-adapted probe gradient. As a linear combination of the original modality gradients, it amounts to a conflict-dependent reweighting that adjusts their relative contributions without introducing a direction outside their linear span. We then apply conditional norm control: with $r_{\max} = \eta \|\mathbf{g}_P\|_2$, where $\eta > 0$ is the norm tolerance ratio, $\mathbf{g}_A$ is rescaled to norm $r_{\max}$ only if $\|\mathbf{g}_A\|_2 > r_{\max}$ and $r_{\max} > \epsilon_g$, and is otherwise left unchanged. The result, denoted $\mathbf{g}_{\text{MCGA}}$, is thus bounded in norm relative to the original gradient sum.

Integration with the total gradient. Let $\mathbf{g}_{\text{total}} = \nabla_{\theta_f} \mathcal{L}_{\text{total}}$ be the gradient from ordinary backpropagation. MCGA replaces the selected modality-gradient sum while retaining remaining contributions:

$$\mathbf{g}_{\text{final}} = \mathbf{g}_{\text{total}} - \mathbf{g}_P + \mathbf{g}_{\text{MCGA}} \tag{14}$$

Parameters outside θ_f retain their ordinary gradients. When no pair satisfies the conflict criterion, Eq. (14) recovers the ordinary total gradient algebraically.

3.4 Training and Inference

We retain the baseline identity, image-prototype, image-text, and prototype-text objectives. Their weighted combination is:

$$\mathcal{L}_{\text{base}} = \sum_{m=0}^{M-1} \omega_m \left[\mathcal{L}^m_{\text{ReID}} + \lambda_{\text{i2p}} \mathcal{L}^m_{\text{i2p}} + \lambda_{\text{i2t}} \mathcal{L}^m_{\text{i2t}} + \lambda_{\text{pt}} \left(\mathcal{L}^m_{\text{p2t}} + \mathcal{L}^m_{\text{t2p}} \right) \right] \tag{15}$$

where λ_{i2p}, λ_{i2t}, and λ_{pt} balance the image-prototype, image-text, and prototype-text alignment losses, respectively. We retain the baseline prototype initialization and memory-update procedures.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \lambda_G(e) \mathcal{L}_{\text{GRASP}} \tag{16}$$

GRASP contributes an additional representation-level penalty, while MCGA acts after gradient computation and is not an additional loss term. Thus, the forward objective remains Eq. (16), while the optimizer receives the modified gradient in Eq. (14) for selected enhancement parameters. At inference, synchronized observations pass through the image encoder and GRACE, and resulting representations are processed by the baseline retrieval head. GRASP and MCGA are not required at inference; GRACE is the only proposed component remaining active in the inference path.

4 EXPERIMENTS

Datasets and Metrics. We evaluate TRIDA on five multi-spectral object ReID benchmarks, comprising two person and three vehicle datasets: MARKET-MM (Wang et al., 2022) and RGBNT201 (Zheng et al., 2021) for persons, and MSVR310 (Zheng et al., 2023b), RGBNT100 (Li et al., 2020), and WMVEID863 (Zheng et al., 2025) for vehicles. MARKET-MM is synthesized from Market-1501 (Zheng et al., 2015) and contains 1,501 identities with 32,668 RGB-NIR-TIR triplets; RGBNT201 contains 4,787 triplets of 201 identities; MSVR310 contains 2,087 triplets of 310 identities; RGBNT100 contains 17,250 triplets of 100 identities; and WMVEID863 contains 4,709 triplets of 863 identities captured by eight cameras. We report mean Average Precision (mAP) and the Cumulative Matching Characteristic (CMC), including Rank-1/5/10 accuracy.

Implementation details. TRIDA is implemented in PyTorch and trained on a single NVIDIA RTX 4090 GPU. We adopt the pre-trained CLIP ViT-B/16 model (Radford et al., 2021) as the backbone and retain the prompt-prototype framework and lightweight modality adapters of ICPL-ReID, with four learnable prompts per modality. Synchronized RGB-NIR-TIR triplets are resized to 256×128 for person datasets and 128×256 for vehicle datasets, and augmented with random horizontal flipping and random erasing (Zhong et al., 2020). We use the Adam optimizer with weight decay 5×10^{-4}, batch size 64, and 4 images per identity. The initial learning rate is 3×10^{-4} for RGBNT201 and MARKET-MM, 3.5×10^{-4} for MSVR310 and WMVEID863, and 1×10^{-4} for RGBNT100. For the proposed components, GRACE partitions channels into $G = 4$ groups, GRASP acts on exchanged CLS features with $\lambda_G^{\max} = 10^{-3}$, a 20-epoch linear warm-up, and $\epsilon = 10^{-6}$. MCGA uses $\tau_c = -0.05$, $\beta_0 = 0.5$, $\rho_p = 1.0$, $\eta = 1.0$, and $\epsilon_g = 10^{-12}$. Models are optimized under a step schedule with a 10-epochs linear warm-up.

4.1 COMPARISON WITH SOTA METHODS

Evaluation on Multi-modal Person ReID. Table 1 compares TRIDA with recent multi-modal person ReID methods on RGBNT201 and MARKET-MM. On RGBNT201, TRIDA achieves 82.6% mAP and 85.3% Rank-1, surpassing the strongest competing results by 1.4% and 0.1%, respectively. Although Signal attains higher Rank-5 and Rank-10 scores, TRIDA remains competitive on both metrics while achieving the best mAP and Rank-1. Moreover, compared to the reproducible ICPL-ReID baseline, TRIDA improves mAP and Rank-1 by 9.6% and 10.5%, respectively, with additional gains of 7.5% and 5.8% in Rank-5 and Rank-10. On MARKET-MM, TRIDA further improves mAP from 85.1% to 88.2% and Rank-1 from 94.7% to 95.5%, while matching the best Rank-5/Rank-10 accuracy at 98.4%/99.1%. This pattern indicates stronger full-ranking retrieval, while the contribution of each component is examined through the ablations below.

Evaluation on Multi-modal Vehicle ReID. Table 2 reports results on MSVR310, RGBNT100, and WMVEID863. On MSVR310, TRIDA reaches 59.4% mAP and 77.0% Rank-1; while STMI obtains a higher mAP (64.8%) and ICPL-ReID gives the best Rank-1 (78.8%), TRIDA improves the reproduced baseline by 2.8% mAP and remains competitive on Rank-1. On RGBNT100, TRIDA achieves 88.5% mAP and the best Rank-1 of 99.2%; its mAP is only 0.6% below STMI while exceeding the reproduced baseline by 2.0%. The improvement is most evident on WMVEID863, where TRIDA is best on all four metrics, obtaining 70.5% mAP, 77.2% Rank-1 (tied with DeMo), 84.3% Rank-5, and 87.0% Rank-10, improving the reproduced ICPL-ReID baseline by 2.3%, 1.9%, 3.3%, and 2.5%, respectively. These gains under severe dazzle coincide with the two stabilizing components: GRASP anchors the exchanged groups against interaction-induced drift, and, as shown in Fig. 8, MCGA arbitrates persistently on the conflict-heavy conditions while remaining nearly inert when modality gradients agree.

Table 1: Comparison with state-of-the-art methods on multi-modal person ReID datasets (%). † ICPL-ReID reproduced under the same setting. Bold: best; underline: second-best.

Methods	Structure	RGBT201				MARKET-MM			
		mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
IEEE (Wang et al., 2022)	CNN	46.4	47.1	58.5	64.2	64.3	83.9	93.0	95.7
TIENet (Yang et al., 2025)	CNN	54.5	54.4	66.3	71.1	67.4	86.1	94.1	96.0
UniCat (Crawford et al., 2023)	ViT	57.0	55.7	–	–	–	–	–	–
EDITOR (Zhang et al., 2024)	ViT	66.7	68.7	82.2	87.9	77.4	90.8	96.8	98.3
RSCNet (Yu et al., 2024)	ViT	68.2	72.5	–	–	–	–	–	–
HTT (Wang et al., 2024b)	ViT	71.1	73.4	83.1	87.3	67.2	81.5	95.8	97.8
TOP-ReID (Wang et al., 2024a)	ViT	72.2	75.2	84.9	89.4	82.0	92.4	97.6	98.6
Signal (Liu et al., 2026)	ViT	80.3	<u>85.2</u>	**91.4**	**93.7**	–	–	–	–
PromptMA (Zhang et al., 2025b)	CLIP	78.4	80.9	87.0	88.9	83.6	93.3	–	–
MambaPro (Wang et al., 2025a)	CLIP	78.9	83.4	89.8	91.9	84.1	92.8	<u>97.7</u>	<u>98.7</u>
DeMo (Wang et al., 2025b)	CLIP	79.7	81.8	89.4	92.5	83.6	93.1	97.5	<u>98.7</u>
IDEA (Wang et al., 2025c)	CLIP	80.2	82.1	90.0	<u>93.3</u>	–	–	–	–
STMI (Xu et al., 2026)	CLIP	<u>81.2</u>	83.4	90.2	91.6	–	–	–	–
ICPL-ReID (Li et al., 2025b)†	CLIP	73.0	74.8	82.8	87.1	<u>85.1</u>	<u>94.7</u>	**98.4**	**99.1**
TRIDA	**CLIP**	**82.6**	**85.3**	<u>90.3</u>	92.9	**88.2**	**95.5**	**98.4**	**99.1**

Table 2: Comparison with state-of-the-art methods on multi-modal vehicle ReID datasets (%).

Methods	Structure	MSVR310		RGBNT100		WMVEID863			
		mAP	R-1	mAP	R-1	mAP	R-1	R-5	R-10
IEEE (Wang et al., 2022)	CNN	21.0	41.0	61.3	87.8	45.9	48.6	64.3	67.9
CCNet (Zheng et al., 2023a)	CNN	36.4	55.2	77.2	96.3	50.3	52.7	69.6	75.1
EDITOR (Zhang et al., 2024)	ViT	39.0	49.3	82.1	96.4	65.6	73.8	80.0	82.3
RSCNet (Yu et al., 2024)	ViT	39.5	49.6	82.3	96.6	–	–	–	–
TOP-ReID (Wang et al., 2024a)	ViT	35.9	44.6	81.2	96.4	67.7	75.3	80.8	83.5
FACENet (Zheng et al., 2025)	ViT	36.2	54.1	81.5	96.9	<u>69.8</u>	77.0	81.0	84.2
Signal (Liu et al., 2026)	ViT	53.6	71.9	86.3	97.6	–	–	–	–
PromptMA (Zhang et al., 2025b)	CLIP	55.2	64.5	85.3	97.4	–	–	–	–
MambaPro (Wang et al., 2025a)	CLIP	47.0	56.5	83.9	94.7	69.5	76.9	80.6	83.8
DeMo (Wang et al., 2025b)	CLIP	49.2	59.8	86.2	97.6	68.8	**77.2**	<u>81.5</u>	83.8
IDEA (Wang et al., 2025c)	CLIP	47.0	62.4	87.2	96.5	–	–	–	–
STMI (Xu et al., 2026)	CLIP	**64.8**	76.1	**89.1**	97.1	–	–	–	–
ICPL-ReID (Li et al., 2025b) †	CLIP	56.6	**78.8**	86.5	<u>98.2</u>	68.2	75.3	81.0	<u>84.5</u>
TRIDA	**CLIP**	<u>59.4</u>	<u>77.0</u>	<u>88.5</u>	**99.2**	**70.5**	**77.2**	**84.3**	**87.0**

4.2 Ablation Study

Tab. 3 reports progressive ablations on RGBNT201 and RGBNT100. On RGBNT201, GRACE produces the largest improvement over the reproduced ICPL-ReID baseline, raising mAP from 73.0% to 80.6% and Rank-1 from 74.8% to 81.5%. GRASP further improves mAP and Rank-1 to 82.3% and 83.7%, while MCGA yields the full TRIDA with 82.6% mAP and 85.3% Rank-1. On RGBNT100, GRACE improves mAP from 86.5% to 87.6%, and GRASP further raises it to 88.5%. MCGA leaves the final performance unchanged at 88.5% mAP and 99.2% Rank-1. This result suggests that the effect of MCGA is conflict-dependent rather than uniformly additive: when modality gradients are already sufficiently compatible, arbitration may introduce little change to the ordinary update.

4.3 Representation and Efficiency Analysis

We use Grad-CAM (Selvaraju et al., 2017) to visualize feature responses across three modalities (Fig. 5). The baseline produces fragmented responses that often extend into irrelevant background regions. Adding GRACE and GRASP progressively concentrates responses on target objects and yields more continuous coverage of identity-discriminative regions. The full TRIDA model generates the most coherent and consistent attention maps across all three spectra while suppressing background clutter. These qualitative results align with the quantitative ablation study.

Table 3: Ablation study of the proposed components on RGBNT201 and RGBNT100.

Method	GRACE	GRASP	MCGA	RGBNT201				RGBNT100			
				mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10
(a) ICPL-ReID† Baseline	✗	✗	✗	73.0	74.8	82.8	87.1	86.5	98.2	98.6	98.7
(b) +GRACE	✓	✗	✗	80.6	81.5	88.9	92.1	87.6	98.9	**99.3**	**99.5**
(c) +GRACE+GRASP	✓	✓	✗	82.3	83.7	90.0	**93.1**	**88.5**	**99.2**	**99.3**	99.4
(d) TRIDA	✓	✓	✓	**82.6**	**85.3**	**90.3**	92.9	**88.5**	**99.2**	**99.3**	99.4

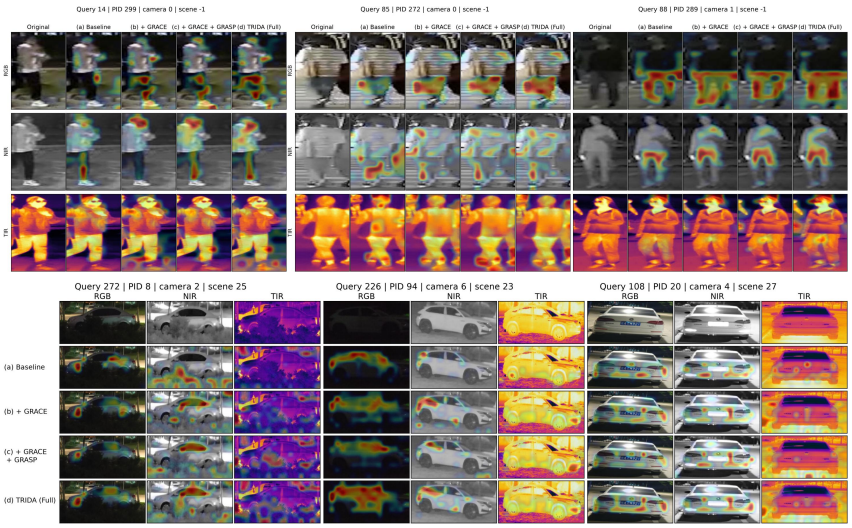

Figure 5: Grad-CAM visualization of the ICPL-ReID baseline and different TRIDA configurations.

To verify that the performance gains are not simply driven by increased model capacity, we analyze parameter efficiency on RGBNT201 in Fig. 6. The ICPL-ReID baseline contains 45.418M trainable parameters. Adding GRACE introduces 3.618M parameters, raising the total to 49.036M (+7.97%), while mAP and Rank-1 increase to 80.6% and 81.5%. On top of GRACE, GRASP and MCGA introduce no additional parameters, yet they further raise mAP to 82.3% and 82.6% and Rank-1 to 83.7% and 85.3%. Overall, TRIDA improves mAP by 9.6% and Rank-1 by 10.5% with only 7.97% more trainable parameters, indicating that the observed gains cannot be explained solely by increased model capacity.

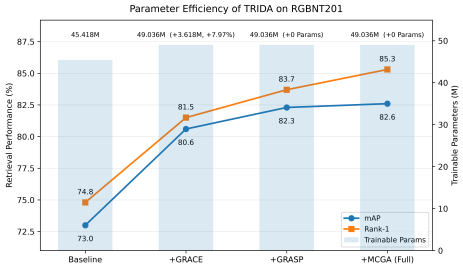

Figure 6: Parameter efficiency of TRIDA.

5 Conclusion and Future Work

We propose TRIDA, a tri-modal residual-aware divergence arbitration framework for multi-spectral object ReID, which addresses three key challenges in RGB-NIR-TIR representation learning: unstructured cross-modal interaction, interaction-induced representation deviation, and conflicting modality-wise optimization directions. GRACE performs structured group-wise feature exchange to exploit complementary spectral information, while GRASP constrains excessive modality-specific deviation from the shared cross-modal center to stabilize exchanged representations. MCGA operates only during training, performing conflict-aware gradient arbitration on the shared interaction parameters to coordinate modality-wise optimization directions. GRACE is the only proposed component retained at inference. Extensive experiments on person and vehicle benchmarks demonstrate the effectiveness and robustness of TRIDA with modest parameter overhead.

AI USE STATEMENT

In this work, generative AI tools were used only as auxiliary support for language polishing, LaTeX and figure/table formatting, preliminary literature screening, and code debugging. They were not used as an independent source of algorithmic ideas, theoretical derivations, or scientific conclusions. The core methodology, mathematical formulations, experimental design, numerical results, and final interpretations were developed, verified, and approved by the authors. All AI-assisted content was carefully reviewed by the authors, who take full responsibility for the final manuscript.

REFERENCES

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. *Advances in Neural Information Processing Systems*, 35:16664–16678, 2022.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pp. 794–803, 2018.

Jennifer Crawford, Haoli Yin, Luke McDermott, and Daniel Cummings. Unicat: Crafting a stronger fusion baseline for multimodal re-identification. In *Proceedings of the UniReps Workshop on Unifying Representations in Neural Models (UniReps) at NeurIPS*, pp. 1–9. NeurIPS Foundation, 2023. URL https://arxiv.org/abs/2310.18812. Extended Abstract. arXiv:2310.18812.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, pp. 1–12, Virtual Event, May 2021.

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. *Advances in Neural Information Processing Systems*, 38:42584–42605, 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 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16796–16807. PMLR, 2025b.

Yujian Feng, Jian Yu, Feng Chen, Yimu Ji, Fei Wu, Shangdong Liu, and Xiao-Yuan Jing. Visible-infrared person re-identification via cross-modality interaction transformer. *IEEE Transactions on Multimedia*, 25:7647–7659, 2023. doi: 10.1109/TMM.2022.3224663.

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 (CVPR)*, Seattle, WA, USA, June 2026. Computer Vision Foundation.

Kongzhu Jiang, Tianzhu Zhang, Xiang Liu, Bingqiao Qian, Yongdong Zhang, and Feng Wu. Cross-modality transformer for visible-infrared person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. doi: 10.1007/978-3-031-19781-9_28.

Chenxin Li, Yifan Liu, Panwang Pan, Hengyu Liu, Xinyu Liu, Wuyang Li, Cheng Wang, Weihao Yu, Yiyang Lin, and Yixuan Yuan. Infobridge: Balanced multimodal integration through conditional dependency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Honolulu, HI, USA, October 2025a. Computer Vision Foundation.

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.

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.

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*, 2025b.

Shihao Li, Chenglong Li, Aihua Zheng, Jin Tang, and Bin Luo. Deep: Decoupled semantic prompt learning, guiding and embedding for multi-spectral object re-identification. *IEEE Transactions on Multimedia*, 28:4731–4746, 2026b.

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.

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 *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2103–2112, 2025.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. *Advances in neural information processing systems*, 34:18878–18890, 2021.

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.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 16428–16446. PMLR, 2022.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8238–8247, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL `http://jmlr.org/papers/v9/vandermaaten08a.html`.

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.

Yuhao Wang, Xuehu Liu, Tianyu Yan, Yang Liu, Aihua Zheng, Pingping Zhang, and Huchuan Lu. Mambapro: Mamba aggregation with synergistic prompt learning for multi-modal object re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 8150–8158, 2025a.

Yuhao Wang, Yang Liu, Aihua Zheng, and Pingping Zhang. 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.

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 (CVPR)*, 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.

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.

Shicai Wei, Chunbo Luo, and Yang Luo. Boosting multimodal learning via disentangled gradient learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22879–22888, 2025.

Ruida Xi. Miss-reid: Delivering robust multi-modality object re-identification despite missing modalities. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, pp. 163372–163396, 2025.

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*, pp. 1–9, 2026.

Xi Yang, Wenjiao Dong, De Cheng, Nannan Wang, and Xinbo Gao. Tienet: A tri-interaction enhancement network for multimodal person reidentification. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):9852–9863, 2025.

Haoli Yin, Jiayao Li, Eva Schiller, Luke McDermott, and Daniel Cummings. Graft: Gradual fusion transformer for multimodal re-identification. *arXiv preprint arXiv:2310.16856*, pp. 1–10, 2023.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

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, 2024.

Haoshuo Zhang, Yufei Bo, Hongwei Zhang, and Meixia Tao. Prompt-based multimodal semantic communication for multi-spectral image segmentation. In *2025 IEEE/CIC International Conference on Communications in China (ICCC)*, pp. 1–6, 2025a. doi: 10.1109/ICCC65529.2025.11148605.

Pingping Zhang, Yuhao Wang, Yang Liu, Zhengzheng Tu, and Huchuan Lu. Magic tokens: Select diverse tokens for multi-modal object re-identification. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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, 2025b.

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.

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*, 100:1–13, 2023a.

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*, 100:274–288, 2023b.

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.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1116–1124, 2015.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International journal of computer vision*, 130(9):2337–2348, 2022.

A Additional Experimental Analyses

A.1 Feature Distribution Visualization

We visualize learned feature distributions of different ablation variants using t-SNE, as shown in Fig. 7. Introducing GRACE improves the compactness of identity-specific features compared with the baseline. With GRASP, intra-identity distributions become further concentrated, indicating that representation deviation is effectively constrained. The full TRIDA model with MCGA yields the most compact and well-structured feature distribution, demonstrating the complementary effectiveness of all three components.

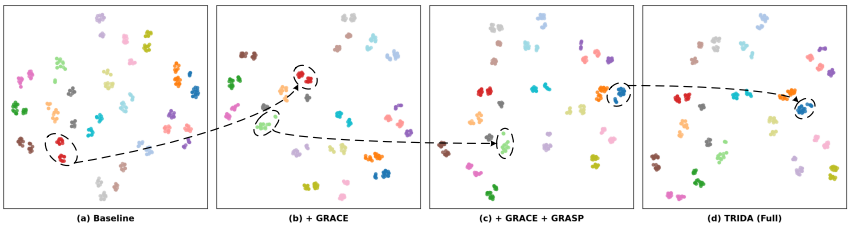

Figure 7: Visualization of the feature distributions with t-SNE (van der Maaten & Hinton, 2008). Different colors denote different identities.

A.2 Gradient Conflict and MCGA Intervention Dynamics

To examine the optimization behavior of MCGA, Fig. 8 visualizes modality-wise gradient conflict and corresponding arbitration strength on MSVR310 and RGBNT201.

On MSVR310, the minimum pairwise cosine similarity decreases to approximately -0.5 in early training and remains negative for most epochs, revealing persistent disagreement between modality-specific optimization directions. The conflict ratio stays around 0.3–0.5, and the effective MCGA blending strength reaches 0.25 under severe disagreement, gradually decreasing as training stabilizes.

In contrast, RGBNT201 exhibits much weaker modality-gradient conflict. After a short transition, the minimum pairwise cosine remains slightly positive, and both conflict ratio and blending strength remain negligible. Thus, MCGA largely preserves the original optimization direction when gradients are compatible, intervening only when significant disagreement is detected, confirming its conflict-adaptive behavior.

A.3 Failure Case Analysis

Fig. 9 presents representative Top-5 retrieval failure cases on RGBNT100, MSVR310, and RGBNT201.

On RGBNT201, failures are mainly caused by strong inter-identity appearance similarity and degraded spectral observations, where different identities share similar clothing patterns, body silhouettes, and infrared responses. On MSVR310, the dominant challenge is fine-grained vehicle ambiguity due to similar body structures and large viewpoint changes. On RGBNT100, several

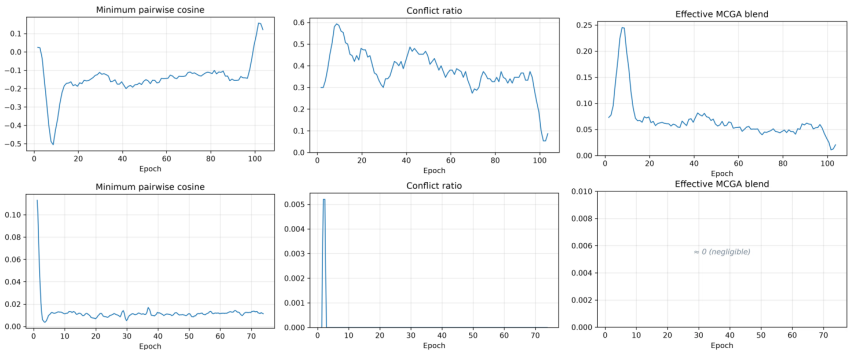

Figure 8: Gradient conflict and MCGA intervention dynamics on MSVR310 and RGBNT201. Top and bottom rows correspond to MSVR310 and RGBNT201, respectively. From left to right: minimum pairwise gradient cosine similarity, conflict ratio, and effective MCGA blending strength.

queries are repeatedly matched to the same incorrect identities, indicating highly ambiguous local neighborhoods in the embedding space.

Overall, remaining errors predominantly occur when discriminative identity evidence is weak or ambiguous across all spectra simultaneously. These cases highlight a future direction: incorporating uncertainty-aware estimation or fine-grained local modeling under severely degraded observations.

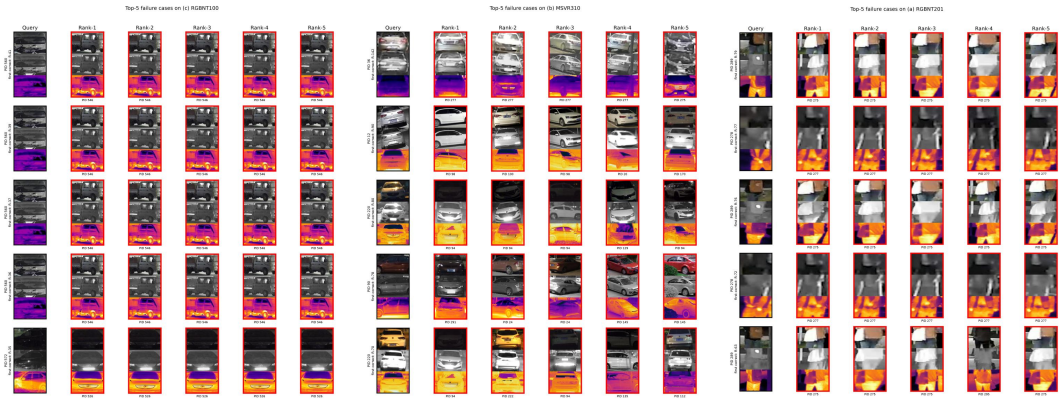

Figure 9: Representative Top-5 retrieval failure cases on (a) RGBNT100, (b) MSVR310, and (c) RGBNT201. Each row shows one query and its five highest-ranked gallery candidates; red borders indicate incorrect matches.