RAPR: RELIABILITY-ADAPTIVE PROMPT ROUTING FOR AERIAL-GROUND PERSON RE-IDENTIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Aerial-ground person re-identification matches pedestrians across drone and ground cameras, whose radically different imaging geometries alter body-part visibility, scale, and context. Consequently, local cues can range from identity-bearing garment details to view-induced clutter and missing regions, making their reliability highly uneven. We present Reliability-Adaptive Prompt Routing (RAPR), a Transformer framework that uses reliability to guide prompt adaptation and local aggregation. RAPR derives a view-reduced identity anchor and dynamically initializes prompts from a shared prompt and an input-conditioned prompt bank. It then selects anchor-consistent local evidence and routes global, reliable-local, and residual view-rejection contexts to refine the prompts. A local refiner separately estimates token identity relevance and view bias to calibrate local memory and guide bidirectional prompt–local interaction. The final descriptor combines global and refined local features. RAPR outperforms multiple state-of-the-art methods across evaluation protocols on several AG-ReID benchmarks. Code and model weights will be released on GitHub upon acceptance.

1 INTRODUCTION

Aerial-ground person re-identification (AG-ReID) retrieves the same pedestrian across aerial and ground cameras, whose different elevations, distances, and fields of view change the evidence available for matching. Aerial images often retain coarse body layout but lose fine identity details; ground images reveal clothing texture and accessories yet are affected by pose, occlusion, and clutter. A cue useful in one view may therefore be weak or misleading in the other. Figure 1 illustrates retrieval in both directions despite changes in scale, pose, visible body regions, and background.

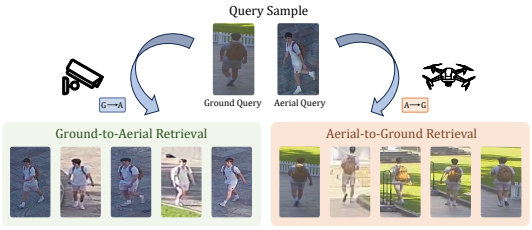

Figure 1: Aerial-ground retrieval in both directions. Queries (top) are matched to opposite-platform galleries (bottom) despite changes in scale, pose, visibility, and background.

Prior AG-ReID work addresses this gap through attribute guidance, elevated-view attention, view decoupling, self-calibrated prompts, and semantic alignment Nguyen et al. (2023; 2024); Zhang et al. (2024a); Wang et al. (2025); Zhang et al. (2026). Transformers model long-range relations among body regions Dosovitskiy et al. (2020); He et al. (2021), while prompt learning adapts visual features to the input or domain without rebuilding the encoder Zhou et al. (2022b); Li et al. (2023); Wu et al. (2024).

Yet local tokens differ in their reliability for cross-view matching. Garment patterns, color arrangements, and accessories can carry stable identity evidence, whereas foreshortening, background mixture, shadows, partial visibility, and scale changes can dominate other patches. Reliability also depends on the sample and view: a detail visible on the ground may disappear or be distorted from above, while coarse aerial evidence may not distinguish similar ground appearances. View-decoupled encoders separate identity and view factors but mainly use the view-reduced feature as a global descriptor Zhang et al. (2024a); they do not use both references to decide which patches

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.

update prompts or enter local aggregation. Admitting view-biased patches into the prompt can propagate nuisance through later prompt–local interactions.

We therefore define reliability relative to identity and view: useful evidence should agree with an identity-oriented reference while being weakly explained by the view representation. Before prompt updating, sparse admission keeps unreliable patches out of the prompt state. Afterward, dense calibration recovers partially informative evidence for local aggregation. Both stages apply the same criterion at different resolutions.

We propose *Reliability-Adaptive Prompt Routing* (RAPR), a view-aware encoder-decoder for AG-ReID. A view-decoupled Transformer produces global, view, and patch tokens that establish the references for subsequent decisions. Dynamic initialization selects an input-conditioned prompt prior; reliability-aware routing admits anchor-consistent patches and combines complementary global and local contexts to update it. A view-suppressed local refiner then calibrates token-level evidence before bidirectional aggregation.

Our contributions are threefold: 1) We formulate AG-ReID adaptation as reliability-controlled evidence admission and introduce RAPR, which unifies sparse prompt-time selection with dense descriptor-time calibration under an identity-versus-view criterion; 2) we develop reliability-aware prompt routing, combining input-conditioned prompt initialization, sparse anchor-relative token selection, and adaptive routing of complementary identity and view-suppressed contexts; and 3) we introduce a view-suppressed local refiner that learns token-level identity-relevance and view-bias gates, using them to calibrate local memory and support reliability-guided bidirectional aggregation.

2 Related Work

General ReID. Person ReID spans part-aware representations Quan et al. (2019), Transformer and vision-language models Han et al. (2022); He et al. (2021); Ni et al. (2023); Li et al. (2023); Yan et al. (2023), and prompt or adaptive-token methods that condition on the input, task, modality, or appearance change Ryoo et al. (2021); Rao et al. (2021); Zhou et al. (2022b;a); Wu et al. (2024); He et al. (2024); Zhou et al. (2026); Huang & Liu (2025). Work on corruption and open-domain robustness Zhang et al. (2024c); Liu & Zhang (2025), identity–clothing consistency, cross-modal and sketch transfer, and general object re-identification Pang et al. (2025); Ding et al. (2026); Gong et al. (2024; 2026); Wan et al. (2026b;a); Shen et al. (2025) broadens the setting. AG-ReID additionally requires deciding whether a local cue survives a viewpoint change.

Aerial-Ground ReID. Image benchmarks established AG-ReID Nguyen et al. (2023; 2024), with later work extending it to cross-platform video Zhang et al. (2024b); Nguyen et al. (2025). VDT Zhang et al. (2024a) decouples identity and view to form a view-reduced descriptor, but does not use the two references to regulate local evidence. SeCap Wang et al. (2025) self-calibrates prompts and refines local features, yet does not compare each patch with both identity and view before prompt updating.

SVPR Zheng et al. (2026) refines features with semantic guidance; SD-ReID Wang et al. (2026) uses view-aware diffusion; and ViSA Zhang et al. (2026) aligns aerial-ground semantics while retaining view-specific cues. These approaches emphasize representation learning. RAPR instead compares each patch with identity and view references to gate prompt updates and calibrate local descriptors, retaining useful identity evidence even when it carries view bias.

3 Method

3.1 Overview

RAPR follows one principle: local evidence should influence the representation only when it is identity-consistent and weakly explained by view. To instantiate this principle, the View-Decoupled Identity Anchor (VDIA) prepends a global token and a view token to the patch sequence for joint Transformer encoding. Given an image x, it produces a global token $\boldsymbol{z}^g$ that integrates person-level evidence, a view token $\boldsymbol{z}^v$ trained to absorb aerial-ground nuisance factors, and N patch tokens

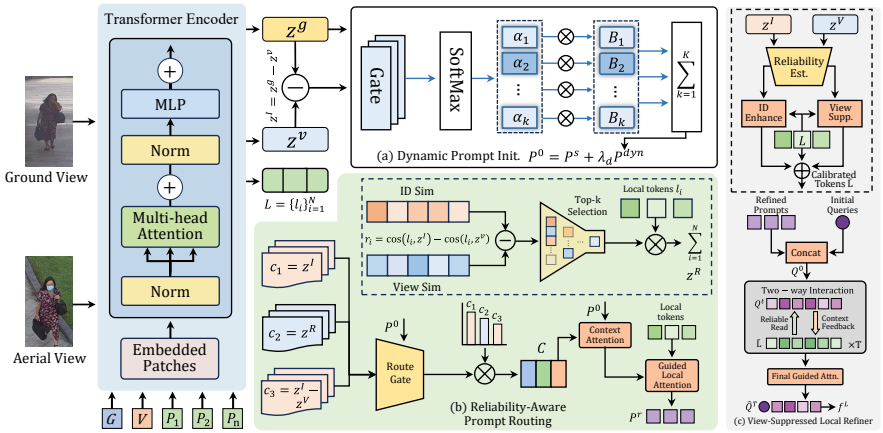

Figure 2: RAPR overview: view-decoupled identity anchoring, reliability-aware prompt routing, and view-suppressed local refinement.

$\boldsymbol{L} = \{\boldsymbol{l}_i\}_{i=1}^{N}$. We construct the view-reduced identity anchor as

$$\boldsymbol{z}^I = \boldsymbol{z}^g - \boldsymbol{z}^v. \tag{1}$$

View supervision and identity–view decorrelation encourage $\boldsymbol{z}^v$ to encode view-specific variation. Its subtraction shifts the global representation away from a learned view direction, yielding an operationally view-reduced anchor rather than an identifiable linear decomposition or guaranteed view invariance. Here, $\boldsymbol{z}^I$ guides prompt adaptation and token assessment, while $\boldsymbol{z}^v$ indicates variation to suppress. VDIA thus supplies paired references to both reliability stages. A global head produces $\boldsymbol{f}^I = H_I(\boldsymbol{z}^I)$, and the local branch refines $\boldsymbol{L}$ with the prompt-guided decoder.

Figure 2 instantiates this chain at two resolutions. RAR applies a nonparametric relative score before prompt refinement and makes a sparse, high-precision admission decision. VSLR acts after prompt context is established and learns dense gates that recover partially informative evidence while suppressing residual view bias. Thus, the stages are not competing reliability estimators: they protect two different decision points in the same evidence path. The global and local branches produce $\boldsymbol{f}^I$ and $\boldsymbol{f}^L$, respectively, which are concatenated for retrieval. RAPR processes each image independently, without paired inputs or cross-view token correspondence.

3.2 Anchor-Conditioned Prompt Routing

Conditioned on the identity anchor, DPI selects an appearance-conditioned prompt prior from the global sample state. RAR then determines which observed evidence may update it and outputs the refined sequence $\boldsymbol{P}^r$ for the local branch.

3.2.1 Dynamic Prompt Initialization

A single shared prompt cannot accommodate the broad range of aerial and ground appearances. We maintain K learnable prompt banks $\{\boldsymbol{B}_k\}_{k=1}^{K}$ and a shared prompt $\boldsymbol{P}^s$, each containing M tokens of dimension D. A gating network predicts image-specific mixture coefficients from the view token and identity anchor:

$$\boldsymbol{\alpha} = \mathrm{softmax}\left(\frac{g_p([\boldsymbol{z}^v; \boldsymbol{z}^I])}{\tau}\right), \tag{2}$$

where g_p is a learnable gating network, τ is a temperature, and $[\cdot;\cdot]$ denotes concatenation. The bank-weighted mixture and its residual injection into the shared prompt are written jointly as

$$\begin{aligned} \boldsymbol{P}^{\mathrm{dyn}} &= \sum_{k=1}^{K} \alpha_k \boldsymbol{B}_k, \\ \boldsymbol{P}^0 &= \boldsymbol{P}^s + \lambda_d \boldsymbol{P}^{\mathrm{dyn}}, \end{aligned} \tag{3}$$

where λ_d is learnable. This retains a shared prior while adapting $\boldsymbol{P}^0$ to the input. The view token remains in the forward path for prompt selection and reliability computation at inference; only its auxiliary classifier, introduced later, is removed. DPI determines how adaptation starts, while RAR decides whether observed evidence is trustworthy enough to alter that state.

3.2.2 RELIABILITY-AWARE ROUTING

Although input-conditioned, $\boldsymbol{P}^0$ remains agnostic to evidence reliability. RAR first forms a compact summary from tokens that are more consistent with identity than with view, then routes it with two global contexts. The anchor-relative score is

$$r_i = \cos(\boldsymbol{l}_i, \boldsymbol{z}^I) - \cos(\boldsymbol{l}_i, \boldsymbol{z}^v). \tag{4}$$

A high score indicates that a token is identity-consistent and weakly view-related. For prompt refinement, we also form a direct view-affinity prior and its logit bias:

$$s_i^V = \cos(\boldsymbol{l}_i, \boldsymbol{z}^v), \qquad b_i^P = -\sigma(\kappa_P) s_i^V,$$

where σ is the sigmoid function and κ_P is a learnable scale. Let $\mathcal{S}$ contain the top 25% of tokens ranked by r_i. We apply a softmax over $\{r_i : i \in \mathcal{S}\}$ to obtain a_i, and set $a_i = 0$ outside $\mathcal{S}$. The reliable local summary is

$$\boldsymbol{z}^R = \sum_{i=1}^{N} a_i \boldsymbol{l}_i. \tag{5}$$

Tokens outside the selected set receive zero weight. This sparse summary prevents unreliable regions from dominating the prompt context. Before context routing, a standard residual self-attention and layer-normalization block lets the M prompt tokens exchange information; for notational economy, its output is still denoted by $\boldsymbol{P}^0$.

RAPR then constructs three candidate contexts:

$$\boldsymbol{c}_1 = \boldsymbol{z}^I, \qquad \boldsymbol{c}_2 = \boldsymbol{z}^R, \qquad \boldsymbol{c}_3 = \boldsymbol{z}^I - \boldsymbol{z}^v. \tag{6}$$

They provide a global fallback, visible fine-grained evidence, and a stronger view-rejection hypothesis. Specifically, $\boldsymbol{c}_3 = \boldsymbol{z}^g - 2\boldsymbol{z}^v$ extrapolates one step farther from the learned view direction. It is a router-selectable context, not a new identity feature or direct retrieval descriptor; the router can suppress it when the first two contexts suffice. A learnable three-way routing network g_r therefore conditions on the average prompt, identity anchor, view token, and local summary:

$$\boldsymbol{\rho} = \mathrm{softmax}\big(g_r([\mathrm{Avg}(\boldsymbol{P}^0); \boldsymbol{z}^I; \boldsymbol{z}^v; \boldsymbol{z}^R])\big). \tag{7}$$

Here, $\mathrm{Avg}(\cdot)$ averages the prompt tokens and $\boldsymbol{\rho}$ is a three-way probability vector. The routed context sequence is $\boldsymbol{C} = [\rho_1\boldsymbol{c}_1; \rho_2\boldsymbol{c}_2; \rho_3\boldsymbol{c}_3]$. The initial prompts first read the routed contexts and are then refined with reliable local tokens:

$$\boldsymbol{P}^c = \mathrm{LN}\big(\boldsymbol{P}^0 + \gamma_c\, \mathrm{Attn}(\boldsymbol{P}^0, \boldsymbol{C}, \boldsymbol{C})\big), \tag{8}$$

$$\hat{\boldsymbol{P}}^r = \mathrm{LN}\big(\boldsymbol{P}^c + \gamma_l\, \mathrm{GAttn}(\boldsymbol{P}^c, \boldsymbol{L}, \boldsymbol{L}; \boldsymbol{a}, \boldsymbol{b}^P)\big),$$

$$\boldsymbol{P}^r = \mathrm{LN}\Big(\hat{\boldsymbol{P}}^r + \mathrm{FFN}(\hat{\boldsymbol{P}}^r)\Big). \tag{9}$$

where γ_c and γ_l are learnable residual scales, Attn denotes standard cross-attention, FFN is a token-wise feed-forward network, and LN denotes layer normalization. The reliability-guided cross-attention is defined as

$$\begin{aligned} &\mathrm{GAttn}(\boldsymbol{X}_q, \boldsymbol{X}_k, \boldsymbol{X}_v; \boldsymbol{\omega}, \boldsymbol{b}) \\ &= \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_h}} + \log(\boldsymbol{\omega} + \epsilon) + \boldsymbol{b}\right)\boldsymbol{V}. \end{aligned} \tag{10}$$

Here, $\boldsymbol{Q} = W_q(\boldsymbol{X}_q)$, $\boldsymbol{K} = W_k(\boldsymbol{X}_k)$, and $\boldsymbol{V} = W_v(\boldsymbol{X}_v)$ are per-head projections, and d_h is the head dimension. The token-wise reliability prior $\boldsymbol{\omega}$ and view bias $\boldsymbol{b}$ are broadcast across heads. Equation 9 uses $(\boldsymbol{X}_q, \boldsymbol{X}_k, \boldsymbol{X}_v) = (\boldsymbol{P}^c, \boldsymbol{L}, \boldsymbol{L})$ and $(\boldsymbol{\omega}, \boldsymbol{b}) = (\boldsymbol{a}, \boldsymbol{b}^P)$. ϵ stabilizes the logarithm. Sparse routing forms $\boldsymbol{z}^R$ exclusively from $\mathcal{S}$ and sharply penalizes unselected keys through $\log\epsilon$ during prompt-to-local attention. VSLR subsequently restores coverage through continuous token weights for correction and aggregation.

3.3 VIEW-SUPPRESSED LOCAL REFINEMENT

VSLR applies the same anchor-conditioned criterion after prompt context is established. It estimates identity relevance and view bias for every token, producing calibrated local memory for descriptor aggregation.

3.3.1 TOKEN RELIABILITY ESTIMATION AND CALIBRATION

Token-level reliability requires high identity relevance and low view bias. Unlike sparse r_i, it is learned for every token. Let $\bar{\boldsymbol{l}}_i$, $\bar{\boldsymbol{z}}^I$, and $\bar{\boldsymbol{z}}^v$ denote layer-normalized local, anchor, and view representations. A lightweight estimator receives them, their element-wise interactions, and cosine similarities:

$$\boldsymbol{\phi}_i = [\bar{\boldsymbol{l}}_i; \bar{\boldsymbol{z}}^I; \bar{\boldsymbol{z}}^v; \bar{\boldsymbol{l}}_i \odot \bar{\boldsymbol{z}}^I; \bar{\boldsymbol{l}}_i \odot \bar{\boldsymbol{z}}^v; \cos(\bar{\boldsymbol{l}}_i, \bar{\boldsymbol{z}}^I); \cos(\bar{\boldsymbol{l}}_i, \bar{\boldsymbol{z}}^v)]. \tag{11}$$

It predicts two latent gates: an identity-conditioned relevance gate p_i^I and a view-conditioned suppression gate p_i^V:

$$[p_i^I, p_i^V] = \sigma(h(\boldsymbol{\phi}_i)). \tag{12}$$

Here, h is the lightweight token estimator. These task-conditioned quantities retain the intended identity-relevance and view-bias roles, but are not calibrated patch probabilities. No patch-level annotation is required; they are optimized end to end through the local descriptor and the image-level objective in Eq. 17. We define token reliability as $u_i = p_i^I(1 - p_i^V)$ and map it to a bounded attention weight $w_i = 0.2 + 0.8u_i$. The calibrated token is

$$\tilde{\boldsymbol{l}}_i = \boldsymbol{l}_i + \eta_I w_i W_I(\boldsymbol{l}_i) - \eta_V p_i^V W_V(\boldsymbol{l}_i), \tag{13}$$

where W_I, W_V, η_I, and η_V are learned. The two terms apply an identity-conditioned residual correction and a subtractive view-conditioned correction, respectively. Calibration adjusts individual token content; the following interaction determines how the routed prompts retrieve and contextualize that evidence.

3.3.2 RELIABILITY-GUIDED BIDIRECTIONAL INTERACTION

The descriptor must now relate calibrated patches to the routed prompts. Let $\tilde{\boldsymbol{L}}^0 = \{\tilde{\boldsymbol{l}}_i\}_{i=1}^N$ and $\boldsymbol{Q}^0 = [\boldsymbol{o}; \boldsymbol{P}^r]$, where $\boldsymbol{o}$ is a learnable output token. Define $\boldsymbol{b}^L = -\sigma(\kappa_L)\boldsymbol{p}^V$ and $\bar{w} = N^{-1}\sum_i w_i$; the forward guided attention therefore uses $(\boldsymbol{\omega}, \boldsymbol{b}) = (\boldsymbol{w}, \boldsymbol{b}^L)$. Omitting positional encodings for clarity, block t performs

$$\begin{aligned}
\bar{\boldsymbol{Q}}^t &= \mathrm{LN}\big(\boldsymbol{Q}^{t-1} + \mathrm{SelfAttn}(\boldsymbol{Q}^{t-1})\big), \\
\hat{\boldsymbol{Q}}^t &= \mathrm{LN}\Big(\bar{\boldsymbol{Q}}^t + \mathrm{GAttn}(\bar{\boldsymbol{Q}}^t, \tilde{\boldsymbol{L}}^{t-1}, \tilde{\boldsymbol{L}}^{t-1}; \boldsymbol{w}, \boldsymbol{b}^L)\Big), \\
\boldsymbol{Q}^t &= \mathrm{LN}\Big(\hat{\boldsymbol{Q}}^t + \mathrm{FFN}(\hat{\boldsymbol{Q}}^t)\Big), \\
\boldsymbol{U}^t &= \mathrm{Attn}(\tilde{\boldsymbol{L}}^{t-1}, \boldsymbol{Q}^t, \boldsymbol{Q}^t), \\
\tilde{\boldsymbol{L}}^t &= \mathrm{LN}\Big(\tilde{\boldsymbol{L}}^{t-1} + [1 + \delta_t(\boldsymbol{w} - \bar{w})] \odot \boldsymbol{U}^t\Big).
\end{aligned} \tag{14}$$

Here, κ_L and δ_t are learned. The forward update lets prompts read reliable local memory while suppressing view-biased keys; the reverse update contextualizes local tokens with the updated prompts and scales their changes by centered reliability. A final guided attention layer maps the output token to $\boldsymbol{f}^L$, complementing $\boldsymbol{f}^I$ with reliable fine-grained evidence.

3.4 TRAINING OBJECTIVE AND INFERENCE

The objective first structures a meaningful identity reference, then makes the global and local descriptors discriminative and identity-consistent. Let y denote the identity label and y^V the binary aerial-versus-non-aerial label; CCTV and wearable observations, when present, belong to the non-aerial class. This binary supervision targets the dominant aerial-to-non-aerial geometric discrepancy rather than attempting to enumerate every camera type. Let C_b be the classifier for identity branch $b \in \{I, L\}$. A view projection produces $\boldsymbol{f}^V = H_V(\boldsymbol{z}^v)$ in the same feature space as $\boldsymbol{f}^I$, while an

Table 1: CARGO comparison across four protocols and their arithmetic mean. **Bold**/underlined mark best/second-best results; ViSA[§] is our reproduction.

Method	Venue	ALL		A→G		G→G		A→A		Average	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
PCB	TPAMI'19	51.00	44.50	34.40	30.40	74.10	67.60	55.00	44.60	53.63	46.78
AGW	TPAMI'21	60.26	53.44	43.57	40.90	81.25	71.66	67.50	56.48	63.15	55.62
SBS	ACM MM'23	50.32	43.09	31.25	29.00	72.31	62.99	67.50	49.73	55.35	46.20
ViT	ICLR'21	61.54	53.54	43.13	40.11	82.14	71.34	80.00	64.47	66.70	57.37
PCL-CLIP	arXiv'23	67.31	60.93	54.37	51.43	84.82	76.00	70.00	60.75	69.13	62.28
CLIP-ReID	AAAI'23	68.27	64.25	55.62	53.83	84.82	**80.80**	75.00	65.42	70.93	66.08
VDT	CVPR'24	64.10	55.20	48.12	42.76	82.14	71.59	**82.50**	66.83	69.22	59.10
DTST	ICME'25	64.42	55.73	50.63	43.39	78.57	72.40	80.00	63.31	68.41	58.71
SeCap	CVPR'25	68.59	60.19	69.43	58.94	86.61	75.42	80.00	68.08	76.16	65.66
SD-ReID	TIP'26	65.06	57.47	53.12	46.44	81.25	74.08	**82.50**	67.70	70.48	61.42
ViSA[§]	CVPR'26	67.63	61.77	63.83	65.51	86.61	80.78	**82.50**	66.23	75.14	68.57
RAPR	Ours	**73.08**	**66.36**	**70.21**	**70.74**	**87.50**	79.69	**82.50**	**70.53**	**78.32**	**71.83**

auxiliary classifier gives $\hat{\boldsymbol{y}}^V = C_V(\boldsymbol{f}^V)$. Identity classification and triplet supervision are applied to both retrieval branches:

$$\mathcal{L}_{\text{reid}} = \sum_{b\in\{I,L\}} \left[\mathcal{L}_{\text{ce}}(C_b(\boldsymbol{f}^b), y) + \mathcal{L}_{\text{tri}}(\boldsymbol{f}^b, y)\right]. \tag{15}$$

View supervision, identity–view decorrelation, and local-to-anchor alignment are defined as

$$\begin{aligned} &\mathcal{L}_{\text{view}} = \mathcal{L}_{\text{ce}}(\hat{\boldsymbol{y}}^V, y^V), \qquad \mathcal{L}_{\text{orth}} = \left|\cos(\boldsymbol{f}^I, \text{sg}(\boldsymbol{f}^V))\right|, \\ &\mathcal{L}_{\text{align}} = 1 - \cos(\boldsymbol{f}^L, \text{sg}(\boldsymbol{f}^I)). \end{aligned} \tag{16}$$

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{reid}} + \lambda_V\left(\mathcal{L}_{\text{view}} + \mathcal{L}_{\text{orth}}\right) + \lambda_A \mathcal{L}_{\text{align}}. \tag{17}$$

The view loss makes $\boldsymbol{f}^V$ encode view factors, while the orthogonality term regularizes their leakage into the identity reference. The one-sided alignment transfers that reference to $\boldsymbol{f}^L$ without allowing noisy local evidence to alter the anchor. The cosine terms are averaged over the mini-batch. λ_V and λ_A balance the terms, and $\text{sg}(\cdot)$ stabilizes the reference branches. At inference, the classification heads C_I, C_L, and C_V are discarded, but $\boldsymbol{z}^v$ remains an internal conditioning signal. Retrieval uses only

$$\boldsymbol{f} = [\boldsymbol{f}^I; \boldsymbol{f}^L]. \tag{18}$$

4 Experiments

4.1 Experimental Settings

Benchmarks and protocols. We evaluate RAPR on three aerial-ground ReID benchmarks. CARGO Zhang et al. (2024a) is our primary benchmark: it contains 108,563 images of 5,000 identities from 13 cameras (five aerial and eight ground). We evaluate it under the all-view, A2G, A2A, and G2G protocols. AG-ReID.v2 Nguyen et al. (2024) contains 100,502 images of 1,615 identities from aerial, CCTV, and wearable cameras; we follow its four official protocols, A2C, A2W, C2A, and W2A. LAGPeR Wang et al. (2025) comprises 63,841 images of 4,231 identities captured by 21 cameras (seven aerial and 14 ground), and we use its A2G, G2A, and G2AG protocols. We follow the official splits in all our experiments. Unless explicitly marked as reproduced, competing results are taken from the cited works. Rank-1 CMC and mAP assess top-ranked matching and overall ranking quality.

Implementation details. RAPR uses an ImageNet-pretrained ViT-B Dosovitskiy et al. (2020); Deng et al. (2009) with 256×128 inputs. We use 64 prompt tokens, retain the top 25% anchor-consistent local tokens in RAR, and employ two VSLR interaction blocks. Training uses SGD for 120 epochs on one RTX 5090, with batch size 64, initial learning rate 8×10^{-3}, a 2,000-iteration warm-up, and cosine decay. We set (λ_V, λ_A) in Eq. 17 to $(10, 0.3)$ on CARGO, $(1, 0.05)$ on AG-ReID.v2, and $(0.01, 0.15)$ on LAGPeR.

Table 2: Comparison on AG-ReID.v2. ViSA[§] reports our reproduction.

Method	Venue	A→C		C→A	
		Rank-1	mAP	Rank-1	mAP
SeCap	CVPR'25	86.88	80.02	85.97	78.15
SD-ReID	TIP'26	87.04	80.61	86.74	79.24
ViSA[§]	CVPR'26	<u>89.26</u>	**83.35**	<u>87.80</u>	<u>81.79</u>
RAPR	Ours	**90.03**	<u>83.03</u>	**89.18**	**82.12**

4.2 Comparison with State-of-the-Art Methods

To assess RAPR across different modeling paradigms and camera configurations, we compare it with representative CNN baselines, Transformer and vision-language models, and recent AG-ReID approaches on CARGO, AG-ReID.v2, and LAGPeR. These include PCB (Sun et al., 2019), AGW (Ye et al., 2021), SBS (He et al., 2023), PCL-CLIP (Li & Gong, 2023), and DTST (Wang & Pishgar, 2025), in addition to the methods discussed in Section 2. These benchmarks jointly cover within-view retrieval and several aerial-ground query directions. CARGO provides the most comprehensive protocol set and therefore serves as our primary benchmark. To preserve space, the main paper reports only representative cross-view protocols on the two auxiliary benchmarks: A2C and C2A on AG-ReID.v2, and A2G and G2A on LAGPeR. Results for their remaining official protocols are provided in the supplementary material. As summarized in Table 1, RAPR achieves the best result under CARGO's ALL protocol, reaching 73.08% Rank-1 and 66.36% mAP. The prior-best ALL mAP is 64.25% from CLIP-ReID, which RAPR improves by 2.11 points. The gain is more pronounced under A→G, where RAPR raises the prior-best mAP from ViSA's 65.51% to 70.74%. This protocol directly exposes the aerial-to-ground appearance gap: fine-grained identity cues are often absent or severely degraded in the aerial query, whereas the gallery contains detailed but differently projected ground observations. The improvement indicates that selectively admitting anchor-consistent local evidence is more effective than treating every patch as equally useful during adaptation.

The same trend extends to the additional benchmarks. On AG-ReID.v2, RAPR attains the highest Rank-1 scores for both reported protocols and the best C→A mAP among the compared methods. Its A→C mAP is 0.32 points below ViSA, despite the substantial change from aerial imagery to fixed-camera observations. On LAGPeR, RAPR exceeds ViSA by 0.56 points in A→G Rank-1 and SD-ReID by 0.21 points in A→G mAP; for G→A, it improves upon SD-ReID by 0.99 Rank-1 points and 0.48 mAP points. These gains show that the routing mechanism generalizes across the evaluated datasets, camera configurations, and query directions. Results for A2W and W2A on AG-ReID.v2 and G2AG on LAGPeR are provided in the supplementary material.

Table 3: Comparison on LAGPeR. ViSA[§] reports our reproduction.

Method	Venue	A→G		G→A	
		Rank-1	mAP	Rank-1	mAP
SeCap	CVPR'25	40.51	29.59	33.95	32.60
SD-ReID	TIP'26	40.15	<u>29.72</u>	<u>34.47</u>	<u>32.86</u>
ViSA[§]	CVPR'26	<u>40.77</u>	29.63	34.04	32.81
RAPR	Ours	**41.33**	**29.93**	**35.46**	**33.34**

4.3 Ablation Studies

To isolate architectural effects from training differences, all variants are initialized from the same ImageNet-pretrained ViT-B backbone and optimized under an identical schedule with identity-classification cross-entropy and triplet supervision. The auxiliary terms in Eq. 17 are enabled only when their associated components are present. The ViT result in Table 1 is reported by prior work, whereas the ViT-B row in Table 4 is retrained under our unified ablation setting and serves as the controlled ReID baseline. We examine the View-Decoupled Identity Anchor (VDIA), Dynamic

Table 4: Component ablation of RAPR on CARGO under four evaluation protocols.

Setting	VDIA	DPI	RAR	VSLR	ALL		A→G		G→G		A→A	
					Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
ViT-B (controlled)					64.31	58.72	59.83	62.92	81.93	72.62	80.00	64.66
+ VDIA	✓				67.88	61.20	65.92	66.10	84.82	75.56	82.50	67.18
+ VDIA + VSLR	✓			✓	69.74	63.02	67.08	67.58	85.71	76.90	82.50	68.18
w/o RAR	✓	✓		✓	70.92	64.38	68.05	68.47	86.61	78.03	82.50	69.15
w/o DPI	✓		✓	✓	71.96	65.35	69.10	69.51	86.61	78.72	82.50	69.68
w/o VSLR	✓	✓	✓		70.66	64.05	67.43	68.02	85.71	77.38	82.50	68.74
Full RAPR	✓	✓	✓	✓	73.08	66.36	70.21	70.74	87.50	79.69	82.50	70.53

Table 5: RAR scoring and sparse-admission ablation on CARGO. $s_i^I = \cos(\boldsymbol{l}_i, \boldsymbol{z}^I)$ and $s_i^V = \cos(\boldsymbol{l}_i, \boldsymbol{z}^v)$; dense admits all tokens. Other components are fixed; bold marks the best results.

Variant	Score	Admission	ALL		A→G	
			Rank-1	mAP	Rank-1	mAP
Uniform	–	Dense	71.10	64.55	68.20	68.60
ID only	s^I	Dense	71.72	65.04	68.86	69.18
Relative	$s^I - s^V$	Dense	72.31	65.57	69.37	69.87
Full RAPR	$s^I - s^V$	Top-25%	**73.08**	**66.36**	**70.21**	**70.74**

Prompt Initialization (DPI), Reliability-Aware Routing (RAR), and View-Suppressed Local Refinement (VSLR). VDIA denotes the complete identity-anchor design, including dual-token encoding, view supervision, identity–view orthogonalization, and construction of the view-reduced identity representation.

Component contributions. Starting from this baseline, VDIA improves the ALL protocol by 3.57 Rank-1 and 2.48 mAP points; the gains rise to 6.09 and 3.18 points on A→G. It also improves G→G and A→A mAP by 2.94 and 2.52 points, respectively. Thus, the anchor improves both severe cross-view matching and within-view discrimination. Adding VSLR further raises ALL and A→G mAP by 1.82 and 1.48 points, indicating that token-level calibration recovers complementary identity evidence while reducing the influence of view-biased regions. Enabling DPI and RAR on top of VDIA and VSLR yields a further 3.34 points in both ALL Rank-1 and mAP. Full RAPR consequently improves over ViT-B by 8.77 Rank-1 and 7.64 mAP points on ALL.

Removing DPI, RAR, or VSLR reduces ALL mAP by 1.01, 1.98, and 2.31 points, with corresponding A→G losses of 1.23, 2.27, and 2.72 points. The DPI drop shows that fixed prompt initialization is insufficient across cameras. RAR admits anchor-consistent context before prompt updating, whereas VSLR suppresses residual view-biased tokens during local refinement. Their distinct drops demonstrate complementary roles and jointly produce the strongest overall result.

Reliability criterion and sparse admission. Table 5 further isolates RAR's scoring and admission rules while holding DPI, context routing, view-affinity bias, and VSLR fixed. Under dense admission, identity-only scoring improves ALL/A→G mAP over uniform token weighting by 0.49/0.58 points, and relative identity-versus-view scoring adds another 0.53/0.69 points. Keeping the relative score but applying Top-25% admission yields a further 0.79/0.87-point gain. Rank-1 follows the same monotonic trend. Full RAR consequently exceeds uniform token weighting by 1.81/2.14 mAP points, validating both relative scoring and sparse admission.

4.4 Model Complexity Analysis

All models in Table 6 are measured on the same CARGO test split with 256×128 inputs, an RTX 5090, and batch size 128; THOP counts one MAC as two FLOPs. Compared with our reproduced

Table 6: Model complexity and inference efficiency on CARGO.

	SeCap	ViSA	RAPR
Venue	CVPR'25	CVPR'26	Ours
GPU	RTX 5090	RTX 5090	RTX 5090
Batch size	128	128	128
Input size	256×128	256×128	256×128
Params (M)	**130.926**	271.778	138.230
GFLOPs	41.760	**37.526**	44.080
Inference time (s)	**57.284**	97.777	61.793
FPS	**890.714**	521.840	825.721
Rank-1 / mAP (%)	68.59 / 60.19	67.63 / 61.77	**73.08 / 66.36**

ViSA, RAPR uses 49.1% fewer parameters, reaches 1.58× the throughput, and improves CARGO ALL Rank-1/mAP by 5.45/4.59 points. Compared with SeCap, it increases both parameter count and GFLOPs by 5.6%, retains 92.7% of its throughput, and gains 4.49/6.17 points. RAPR therefore remains close to SeCap's efficiency while avoiding ViSA's parameter cost.

4.5 Visualizations

Distance distributions. Figure 3 compares pairwise cosine distances on CARGO A→G. RAPR shifts matched pairs toward smaller distances and increases the mean intra/inter-class gap from 0.542 to 0.556, reducing overlap at ambiguous distances. This pattern is consistent with suppressing view-dominated evidence; other protocols appear in the supplementary material.

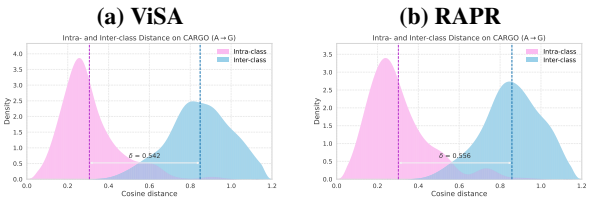

Figure 3: Intra-/inter-class cosine-distance distributions on CARGO A→G; dashed lines mark the means.

Retrieval results. Figure 4 compares Rank-5 lists on CARGO A→G and LAGPeR G→A. RAPR places a positive first for both CARGO queries and retrieves five positives for each LAGPeR query, whereas ViSA misses Rank-1 on CARGO and returns no positives on LAGPeR. This contrast indicates better retention of clothing and silhouette cues under viewpoint and resolution changes.



Figure 4: Rank-5 retrieval on CARGO A→G and LAGPeR G→A (green: correct; red: incorrect).

5 Conclusion

We presented RAPR, which treats aerial-ground person re-identification as reliability-controlled evidence admission. Guided by a view-reduced identity anchor, RAPR couples sparse prompt-time selection with dense view-suppressed local aggregation under a shared identity-versus-view criterion. Results and targeted analyses on CARGO, AG-ReID.v2, and LAGPeR validate relative reliability and sparse admission across view gaps. Without paired cross-view inputs or patch-level annotations, RAPR retains practical inference efficiency. Future work will learn reliability from unlabeled or continuously varying viewpoints to reduce dependence on coarse view labels.

AI USE STATEMENT

Generative AI tools were used to polish the language of the manuscript.

REFERENCES

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Yongkang Ding, Zi Ye, Ivonne Xu, Shuangquan Lyu, and Liyan Zhang. Multi-scale adaptive clustering and local consistency learning for unsupervised clothing-changing person re-identification. *IEEE Transactions on Information Forensics and Security*, 2026.

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. *arXiv preprint arXiv:2010.11929*, 2020.

Yunpeng Gong, Zhun Zhong, Yansong Qu, Zhiming Luo, Rongrong Ji, and Min Jiang. Cross-modality perturbation synergy attack for person re-identification. *Advances in Neural Information Processing Systems*, 37:23352–23377, 2024.

Yunpeng Gong, Yongjie Hou, Jiangming Shi, Kim Long Diep, and Min Jiang. A theory-inspired framework for few-shot cross-modal sketch person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4284–4292, 2026.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 9664–9667, 2023.

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.

Weizhen He, Yiheng Deng, Shixiang Tang, Qihao Chen, Qingsong Xie, Yizhou Wang, Lei Bai, Feng Zhu, Rui Zhao, Wanli Ouyang, et al. Instruct-reid: A multi-purpose person re-identification task with instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17521–17531, 2024.

Zhizhong Huang and Xiaoming Liu. Generalizable object re-identification via visual in-context prompting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22539–22550, 2025.

Jiachen Li and Xiaojin Gong. Prototypical contrastive learning-based clip fine-tuning for object re-identification. *arXiv preprint arXiv:2310.17218*, 2023.

Siyuan Li, Li Sun, and Qian 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.

Zhipu Liu and Lei Zhang. Multi-model synergy perception for open-world person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(11):10891–10905, 2025. doi: 10.1109/TCSVT.2025.3577279.

Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Aerial-ground person re-id. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 2585–2590. IEEE, 2023.

Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Ag-reid.v2: Bridging aerial and ground views for person re-identification. *IEEE Transactions on Information Forensics and Security*, 19:2896–2908, 2024.

Huy Nguyen, Kien Nguyen, Akila Pemasiri, Feng Liu, Sridha Sridharan, and Clinton Fookes. Ag-vpreid: A challenging large-scale benchmark for aerial-ground video-based person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1241–1251, 2025.

Hao Ni, Yuke Li, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Part-aware transformer for generalizable person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11280–11289, 2023.

Zhiqi Pang, Junjie Wang, Lingling Zhao, and Chunyu Wang. Identity-clothing similarity modeling for unsupervised clothing change person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19251–19260, 2025.

Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. Auto-reid: Searching for a part-aware convnet for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3750–3759, 2019.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Michael S Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *Advances in Neural Information Processing Systems*, 34, 2021.

Fei Shen, Xiaoyu Du, Liyan Zhang, Xiangbo Shu, and Jinhui Tang. Triplet contrastive representation learning for unsupervised vehicle re-identification. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(11):1–23, 2025.

Yifan Sun, Liang Zheng, Yali Li, Yi Yang, Qi Tian, and Shengjin Wang. Learning part-based convolutional features for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):902–917, 2019.

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. *Advances in Neural Information Processing Systems*, 38:80232–80255, 2026a.

Xixi Wan, Aihua Zheng, Zi Wang, Bo Jiang, Jin Tang, and Jixin Ma. Reliable multi-modal object re-identification via modality-aware graph reasoning. *IEEE Transactions on Information Forensics and Security*, 2026b.

Shining Wang, Yunlong Wang, Ruiqi Wu, Bingliang Jiao, Wenxuan Wang, and Peng Wang. Secap: Self-calibrating and adaptive prompts for cross-view person re-identification in aerial-ground networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22119–22128, 2025.

Yuhai Wang and Maryam Pishgar. Dynamic token selective transformer for aerial-ground person re-identification. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

Yuhao Wang, Xiang Hu, Lixin Wang, Pingping Zhang, and Huchuan Lu. Sd-reid: View-aware stable diffusion for aerial-ground person re-identification. *IEEE Transactions on Image Processing*, 35: 5686–5697, 2026. doi: 10.1109/TIP.2026.3694135.

Ruiqi Wu, Bingliang Jiao, Wenxuan Wang, Meng Liu, and Peng Wang. Enhancing visible-infrared person re-identification with modality- and instance-aware visual prompt learning. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 579–588, 2024.

Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing*, 32:6032–6046, 2023.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

Quan Zhang, Lei Wang, Vishal M Patel, Xiaohua Xie, and Jianhuang Lai. View-decoupled transformer for person re-identification under aerial-ground camera network. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22000–22009. IEEE, 2024a.

Quan Zhang, Zeqiang Cai, Peiming Zhao, Jingze Wu, Cailun Wu, Hongbo Chen, and Jianhuang Lai. View-aware semantic alignment for aerial-ground person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4383–4392, 2026.

Shizhou Zhang, Wenlong Luo, De Cheng, Qingchun Yang, Lingyan Ran, Yinghui Xing, and Yanning Zhang. Cross-platform video person reid: A new benchmark dataset and adaptation approach. In *European Conference on Computer Vision*, pp. 270–287. Springer, 2024b.

Xinyu Zhang, Peng Zhang, and Caifeng Shan. Corruption-invariant person re-identification via coarse-to-fine feature alignment. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2):1084–1097, 2024c.

Aihua Zheng, Hao Xie, Xixi Wan, Zi Wang, Shihao Li, Jin Tang, and Bin Luo. Semantic-driven visual progressive refinement for aerial-ground person reid: A challenging large-scale benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13360–13368, 2026. doi: 10.1609/aaai.v40i16.38339.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16816–16825, 2022a.

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, 2022b.

Linhan Zhou, Shuang Li, Neng Dong, Yonghang Tai, Yafei Zhang, and Huafeng Li. Hierarchical prompt learning for image-and text-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13728–13736, 2026.

A Appendix

A.1 Datasets and Evaluation Protocols

CARGO. CARGO is our primary benchmark. It contains 108,563 images of 5,000 identities captured by five aerial and eight ground cameras. The substantial altitude and resolution differences between these cameras produce severe changes in body scale, visible parts, and background composition. We follow four official protocols. ALL jointly evaluates the available aerial and ground views; A→G uses aerial queries and a ground gallery; G→G and A→A measure within-view retrieval for the ground and aerial cameras, respectively. The latter two protocols provide useful controls for determining whether improved cross-view invariance compromises discrimination within either view.

AG-ReID.v2. AG-ReID.v2 comprises 100,502 images of 1,615 identities observed by aerial cameras, fixed CCTV cameras, and wearable devices. Unlike a two-view benchmark, it contains two distinct ground modalities: CCTV imagery retains a relatively stable scene geometry, whereas wearable cameras introduce stronger ego-motion and framing variation. Its A→C, A→W, C→A, and W→A protocols evaluate both directions for each aerial-ground pairing. These directional tests are not interchangeable because changing the query view also changes the gallery composition and ranking difficulty. We therefore report all four protocols and their arithmetic average in Table 7.

LAGPeR. LAGPeR comprises 63,841 images of 4,231 identities captured by seven aerial and 14 ground cameras. It provides A→G and G→A for bidirectional cross-view retrieval. The G→(A+G) protocol further places aerial and ground images in the same gallery. This heterogeneous gallery is more than a simple union of the preceding settings: same-view candidates can dominate the early ranking, while true matches from the other view retain large scale and viewpoint discrepancies. Table 8 reports all three protocols and their arithmetic average.

Evaluation protocols. For every benchmark, we use its official training, query, and gallery partitions without altering the identity split. We report Rank-1 cumulative matching characteristic accuracy and mean Average Precision (mAP). Rank-1 measures whether the first retrieved image contains the correct identity, whereas mAP summarizes precision over all valid matches in the ranked list and is therefore more sensitive to difficult positives. All values are percentages. The Average columns in the supplementary tables are arithmetic means over the listed protocols and are included to summarize directional stability, not to replace protocol-specific comparisons.

A.2 Extended Comparison with State-of-the-Art Methods

The main paper reports the complete CARGO comparison but retains only representative AG-ReID.v2 and LAGPeR protocols because of the page limit. The supplementary results restore every official direction, allowing performance to be examined under distinct ground-camera types, reversed query-gallery roles, and a heterogeneous gallery. The compared methods cover conventional CNN and Transformer baselines, vision-language ReID models, and recent AG-ReID-specific approaches. Unless a row is explicitly marked as our reproduction, its values are taken from the cited work. Reproduced rows use the same dataset protocols as RAPR and are marked to distinguish them from author-reported results.

A.2.1 Results on AG-ReID.v2

The complete comparison in Table 7 shows that RAPR achieves the best average Rank-1 and mAP, reaching 90.08% and 83.39%, respectively. It improves the strongest previous average Rank-1 result, obtained by V2E, by 0.36 points and exceeds our reproduced ViSA by 0.08 points in average mAP. The latter margin is modest, but it is accompanied by leading results in six of the eight protocol-specific metric columns. This distinction is important: the average improvement does not arise from an isolated gain under one favorable query direction. Since the four protocols vary in ground-camera type, query direction, and gallery composition, we consider their individual results together with the average rather than treating the benchmark as a single pooled retrieval task. On A→C, RAPR improves Rank-1 from ViSA's 89.26% to 90.03% while retaining the second-best mAP. Under A→W, V2E remains strongest in Rank-1, whereas RAPR achieves the highest mAP of 85.78%.

Table 7: Performance comparison on AG-ReID.v2. Average is computed over the four protocols; C, W, and A denote CCTV, wearable, and aerial views. Best and second-best results are **bold** and underlined; ViSA[§] is our reproduction.

Method	Venue	A→C		A→W		C→A		W→A		Average	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Swin	ICCV'21	68.76	57.66	68.49	56.15	68.80	57.70	64.40	53.90	67.61	56.35
HRNet-18	TPAMI'20	75.21	65.07	76.26	66.17	76.25	66.16	76.25	66.17	75.99	65.89
SwinV2	CVPR'22	76.44	66.09	80.08	69.09	77.11	62.14	74.53	65.61	77.04	65.73
MGN (R50)	ACM MM'18	82.09	70.17	88.14	78.66	84.21	72.41	84.06	73.73	84.63	73.74
BoT (R50)	CVPR'19	80.73	71.49	86.06	75.98	79.46	69.67	82.69	72.41	82.24	72.39
SBS (R50)	ACM MM'23	81.96	72.04	88.14	78.94	84.10	73.89	84.66	75.01	84.72	74.97
BoT (ViT)	CVPR'19	85.40	77.03	89.77	80.48	84.65	75.90	84.65	75.90	86.12	77.33
ViT	ICLR'21	85.40	77.03	89.77	80.48	84.65	75.90	84.27	76.59	86.02	77.50
TransReID	ICCV'21	88.00	81.40	90.40	84.50	87.60	80.10	87.70	81.10	88.43	81.78
FusionReID	T-ITS'25	86.70	80.70	89.70	84.20	87.90	80.00	86.50	80.90	87.70	81.45
CLIP-ReID	AAAI'23	85.36	79.79	89.14	84.23	85.64	79.08	86.50	79.55	86.66	80.66
PCL-CLIP	arXiv'23	79.80	72.20	87.14	77.70	81.12	72.40	84.19	73.89	83.06	74.05
VDT	CVPR'24	86.46	79.13	90.00	82.21	86.14	78.12	85.26	78.52	86.97	79.50
V2E (ViT)	TIFS'24	88.77	80.72	**93.62**	84.85	87.86	78.51	88.61	80.11	89.72	81.05
SeCap	CVPR'25	86.88	80.02	90.06	82.89	85.97	78.15	85.17	78.54	87.02	79.90
SD-ReID	TIP'26	87.04	80.61	90.86	84.06	86.74	79.24	86.79	80.12	87.86	81.01
ViSA[§]	CVPR'26	89.26	**83.35**	91.59	85.67	87.80	81.79	89.06	82.41	89.43	83.31
RAPR	Ours	**90.03**	83.03	91.85	**85.78**	**89.18**	**82.12**	**89.27**	**82.64**	**90.08**	**83.39**

RAPR obtains both best metrics in the two reverse directions. Relative to the second-best entries, it improves C→A by 1.28/0.33 points and W→A by 0.21/0.23 points in Rank-1/mAP. This consistency across fixed CCTV and moving wearable cameras indicates that reliability-aware routing is not specialized to one ground-camera geometry. The comparison across method families provides further context: conventional CNN baselines remain below 75% average mAP, TransReID reaches 81.78% without an AG-ReID-specific design, and recent view-aware methods narrow the gap further. ViSA is the strongest competitor in average mAP, while V2E retains a 1.77-point Rank-1 lead on A→W. RAPR nevertheless achieves higher mAP on that protocol and leads both metrics when the aerial view forms the gallery. The four directions therefore support a balanced improvement in ranking complete match sets rather than an advantage concentrated in one metric or query direction.

A.2.2 Results on LAGPeR

Table 8 extends the main-paper comparison to all LAGPeR protocols. RAPR ranks first in every Rank-1 and mAP column and obtains the best average of 33.44% Rank-1 and 27.41% mAP. Compared with the strongest competing entry for each metric, the gains are 0.56/0.21 points on A→G, 0.99/0.48 points on G→A, and 0.33/0.08 points on G→(A+G) in Rank-1/mAP. At the aggregate level, RAPR improves average Rank-1 by 0.77 points over our reproduced ViSA and average mAP by 0.26 points over SD-ReID.

The bidirectional protocols expose a clear asymmetry: all methods obtain higher Rank-1 for aerial queries than for ground queries, whereas the mAP gap is considerably smaller.

This suggests that a ground query must compete against more visually ambiguous aerial candidates at the top of the ranking, even when multiple correct matches remain recoverable. RAPR improves both directions rather than trading one for the other. The mixed-gallery protocol is the hardest, with substantially lower scores for every method. Its same-view ground candidates and cross-view aerial candidates provide different levels of identity evidence within one ranked list. RAPR's positive gain under this setting, although smaller than under G→A, is therefore complementary evidence that the router can regulate heterogeneous local evidence instead of relying on a fixed preference for either aerial or ground appearance.

Table 8: Performance comparison on LAGPeR. Average is the arithmetic mean over the three protocols. **Bold** and underlined entries denote the best and second-best results, respectively. A→G uses aerial images as queries, G→A uses ground images as queries, and G→(A+G) uses aerial and ground images in the gallery. CLIP-ReID* uses OLP and SIE; MIP† is a re-implementation for AG-ReID; ViSA§ reports our reproduction.

Method	Venue	A→G		G→A		G→(A+G)		Average	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
ViT	ICLR'21	38.67	27.25	32.04	30.69	18.88	15.31	29.86	24.42
TransReID	ICCV'21	38.80	28.80	33.00	32.10	22.90	18.80	31.57	26.57
CLIP-ReID	AAAI'23	24.40	17.60	21.30	20.80	12.30	10.20	19.33	16.20
CLIP-ReID*	AAAI'23	23.10	17.50	20.00	20.30	9.00	8.40	17.37	15.40
MIP†	ICMR'24	39.30	29.30	33.90	32.60	21.00	17.30	31.40	26.40
VDT	CVPR'24	40.15	28.97	33.55	31.98	19.50	16.45	31.07	25.80
SeCap	CVPR'25	40.51	29.59	33.95	32.60	22.23	17.97	32.23	26.72
SD-ReID	TIP'26	40.15	<u>29.72</u>	<u>34.47</u>	<u>32.86</u>	22.85	<u>18.88</u>	32.49	<u>27.15</u>
ViSA§	CVPR'26	<u>40.77</u>	29.63	34.04	32.81	<u>23.21</u>	18.37	<u>32.67</u>	26.94
RAPR	Ours	**41.33**	**29.93**	**35.46**	**33.34**	**23.54**	**18.96**	**33.44**	**27.41**

Table 9: Ablation of candidate contexts and adaptive routing on CARGO and LAGPeR. $\boldsymbol{c}_1$, $\boldsymbol{c}_2$, and $\boldsymbol{c}_3$ denote the identity anchor, reliable local summary, and strongly view-suppressed context, respectively. Best and second-best results are bold and underlined.

Variant	$\boldsymbol{c}_1$	$\boldsymbol{c}_2$	$\boldsymbol{c}_3$	Fusion	CARGO				LAGPeR			
					ALL		A→G		A→G		G→A	
					Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
VDIA only	✓			–	67.88	61.20	65.92	66.10	39.92	28.67	33.96	31.92
Anchor + reliable context	✓	✓		Equal	72.06	65.23	69.18	69.54	40.47	29.11	34.51	32.43
Three contexts, equal	✓	✓	✓	Equal	<u>72.52</u>	<u>65.78</u>	<u>69.71</u>	<u>70.16</u>	<u>40.88</u>	<u>29.52</u>	<u>35.02</u>	<u>32.91</u>
Full routing	✓	✓	✓	Learned	**73.08**	**66.36**	**70.21**	**70.74**	**41.33**	**29.93**	**35.46**	**33.34**

A.3 Additional Ablation Studies

A.3.1 Candidate Contexts and Adaptive Routing

We use a controlled progression to isolate the three candidate contexts and the learned router. The first row is the VDIA-only configuration reported in the main-paper component ablation. Starting from this identity anchor, the second row adds the reliable local summary under parameter-free equal fusion, and the third further introduces the strongly view-suppressed context while preserving the same fusion rule. The "Equal" assigns the available contexts identical, input-independent weights, whereas "Learned" uses the sample-conditioned routing probabilities $\boldsymbol{\rho}$ defined in the main paper. The final comparison therefore changes only fixed equal fusion to sample-adaptive routing. The successive differences therefore measure the contributions of $\boldsymbol{c}_2$, $\boldsymbol{c}_3$, and learned routing, respectively. CARGO and the two bidirectional LAGPeR protocols test whether these effects persist across datasets and query directions.

The results exhibit the same progressive trend on both benchmarks. Adding the reliable local summary consistently improves the identity anchor, and introducing the view-suppressed context yields a further gain under the unchanged equal-fusion rule. Replacing equal fusion with learned routing produces the best result in every column. On CARGO, this final step improves ALL Rank-1/mAP by 0.56/0.58 points and A→G by 0.50/0.58 points. The corresponding gains on LAGPeR are 0.45/0.41 points for A→G and 0.44/0.43 points for G→A, indicating that adaptive context selection remains useful under reversed query-gallery roles.

Table 10: Sensitivity of RAR to the token admission ratio on CARGO. The relative score $s_i^I - s_i^V$ is used for every setting. Best and second-best results are bold and underlined.

Admission ratio	ALL		A→G	
	Rank-1	mAP	Rank-1	mAP
10%	72.24	65.50	69.41	69.95
25%	**73.08**	**66.36**	**70.21**	**70.74**
50%	<u>72.79</u>	<u>66.05</u>	<u>69.93</u>	<u>70.43</u>
75%	72.08	65.31	69.12	69.56
100%	72.31	65.57	69.37	69.87

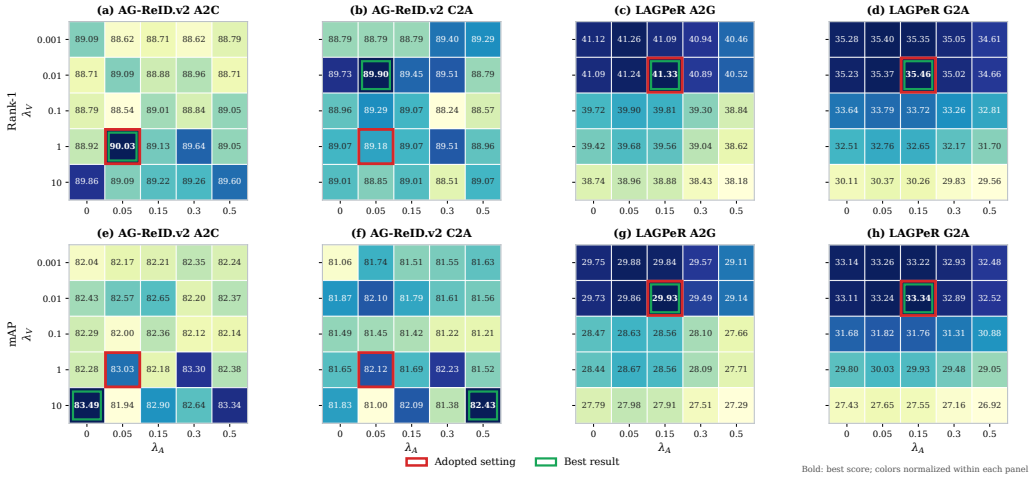

Figure 5: Joint sensitivity to λ_V and λ_A on AG-ReID.v2 and LAGPeR. The top and bottom rows report Rank-1 and mAP, respectively. Red boxes mark the adopted settings, green boxes mark the best result in each panel, and all values are percentages.

A.3.2 Sensitivity to the Admission Ratio

We vary the fraction of patch tokens admitted by RAR while keeping the relative identity-versus-view score, prompt initialization, context routing, VSLR, and training configuration unchanged. Hence, only the number of tokens allowed to refine the prompt changes across rows. Results on ALL and A→G jointly test the aggregate setting and the most relevant cross-view retrieval direction.

The 25% ratio performs best on both protocols. Relative to admitting only 10% of the tokens, it improves ALL Rank-1/mAP by 0.84/0.86 points and A→G by 0.80/0.79 points, showing that an overly restrictive mask discards useful identity evidence. Expanding the ratio from 25% to 50% causes only a 0.28–0.31 point decrease, which indicates a stable neighborhood around the selected setting. Ratios above 50% remain consistently below the 25% setting, although dense admission produces a modest rebound over 75%. Relative to 25%, admitting all tokens reduces ALL by 0.77/0.79 points and A→G by 0.84/0.87 points. This pattern indicates that dense admission reintroduces view-dominated responses and weakens the purpose of reliability filtering. The agreement between the two protocols supports 25% as a practical balance between evidence coverage and selectivity rather than an isolated, sharply tuned optimum.

A.3.3 Hyperparameter Sensitivity

The coefficients λ_V and λ_A weight view supervision and local-to-anchor alignment in the training objective; they are not inference-time parameters. We jointly evaluate $\lambda_V \in \{0.001, 0.01, 0.1, 1, 10\}$ and $\lambda_A \in \{0, 0.05, 0.15, 0.3, 0.5\}$. Each grid cell corresponds to a separately trained model for which both coefficients remain fixed throughout training. Figure 5 reports Rank-1 and mAP on the

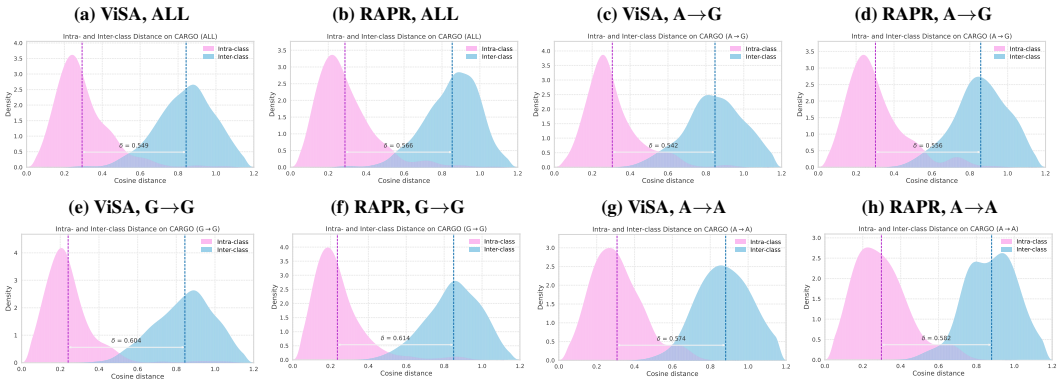

Figure 6: Pairwise cosine-distance distributions of ViSA and RAPR on CARGO under the ALL, A→G, G→G, and A→A protocols.

A→C and C→A protocols of AG-ReID.v2 and the A→G and G→A protocols of LAGPeR. The two-dimensional sweep evaluates both the individual effects of the losses and their interaction. Neither benchmark provides an official validation split; we therefore adopt one dataset-level configuration for all of its evaluation protocols and perform no protocol-specific tuning.

On AG-ReID.v2, the adopted setting $(\lambda_V, \lambda_A) = (1, 0.05)$ provides a favorable joint operating point. It obtains the highest aerial-to-ground Rank-1 of 90.03%, while remaining close to the individual optima in the reverse direction and in mAP. Although other combinations attain a higher value in a single panel, their preferred regions are inconsistent across directions and metrics. For example, stronger view supervision improves the best mAP values but does not produce the best Rank-1 in both directions. Selecting a common setting therefore avoids tuning the model to one particular evaluation view.

LAGPeR exhibits a clearer optimum at $(0.01, 0.15)$, which achieves 41.33/29.93% Rank-1/mAP for aerial-to-ground retrieval and 35.46/33.34% for the reverse direction. Increasing λ_V beyond 0.01 consistently degrades both metrics, indicating that excessive view supervision can preserve view-specific variation at the expense of identity discrimination on this more challenging benchmark. Very large λ_A values also reduce performance, whereas the moderate alignment weight of 0.15 gives the best balance. Overall, the broad neighborhoods around the adopted settings remain competitive, while the dataset-specific optima support using different loss weights for the two benchmarks.

B Additional Visualization Analysis

B.1 Feature-Distance Distributions

We further compare the pairwise cosine-distance distributions of ViSA and RAPR on all four CARGO protocols in Figure 6. Pink and blue densities denote intra-class and inter-class pairs, respectively. Dashed lines indicate their mean distances, and δ denotes the gap between the two means. The A→G pair is reported in the main paper; the full set is provided here for completeness.

RAPR enlarges the mean separation under every protocol. Compared with ViSA, δ increases from 0.549 to 0.566 on ALL, from 0.542 to 0.556 on A→G, from 0.604 to 0.614 on G→G, and from 0.574 to 0.582 on A→A. The A→G setting has the smallest gap and the most visible overlap for both methods, which is consistent with the stronger appearance mismatch in cross-view retrieval. In contrast, the two within-view protocols show larger separations, especially G→G, because camera geometry and visible body regions are more consistent.

The improvement is nevertheless present in both cross-view and within-view settings. RAPR shifts matching pairs toward smaller distances while retaining nonmatching pairs at larger distances, reducing the ambiguous region where their density tails overlap. The gain on ALL further shows that this separation is maintained when aerial and ground samples coexist in one evaluation pool. Therefore, the performance gains do not result from improving one favorable camera configuration

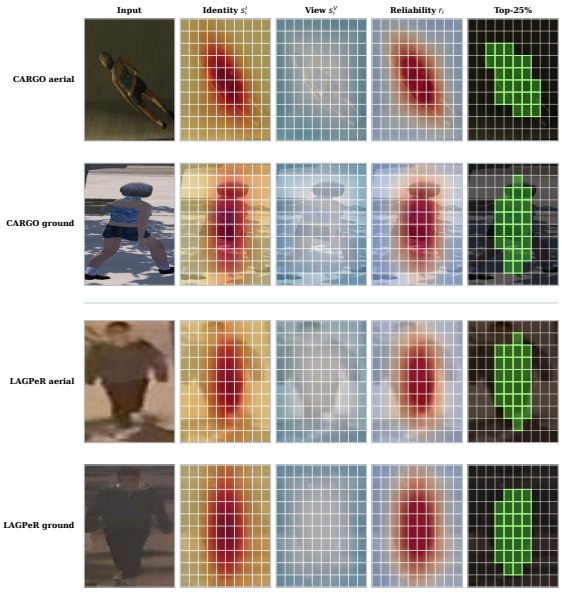

Figure 7: Token reliability on CARGO and LAGPeR: input, identity similarity s_i^I, view similarity s_i^V, relative reliability $r_i = s_i^I - s_i^V$, and RAR's top-25% tokens (left to right).

at the expense of another; the learned descriptor provides more consistent identity separation across all four CARGO protocols.

B.2 Reliability Heatmaps

Figure 7 shows how RAR distinguishes identity-bearing tokens from view-specific local evidence. Identity similarity responds broadly to pedestrian regions, while view similarity highlights descriptors aligned with the view token. Their overlap shows that a body token can carry both identity and view signals. RAR subtracts the view response to sharpen relative reliability and ranks the scores for sparse admission. The resulting maps differ from identity maps or generic foreground saliency: regions responsive to both references are attenuated, potentially changing their rank even on the pedestrian. The final mask prioritizes identity-dominant evidence without uniformly covering the whole body.

The selected regions adapt to image content and capture view. In CARGO, RAR follows the oblique aerial body's elongated support and the ground subject's visible torso and legs despite pose and scale changes. Under stronger LAGPeR blur, admitted tokens stay centered on the pedestrian silhouette and exclude most surroundings. The paired references thus produce sample-adaptive evidence rather than a fixed spatial pattern; consistent behavior across both datasets and views shows that the criterion is not tied to one camera geometry. These observations complement the controlled admission-ratio study in Table 10.