BOUNDED CONTEXTUAL REFINEMENT WITH NEIGHBORHOOD SUPPORT FOR TEXT-AERIAL PERSON RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

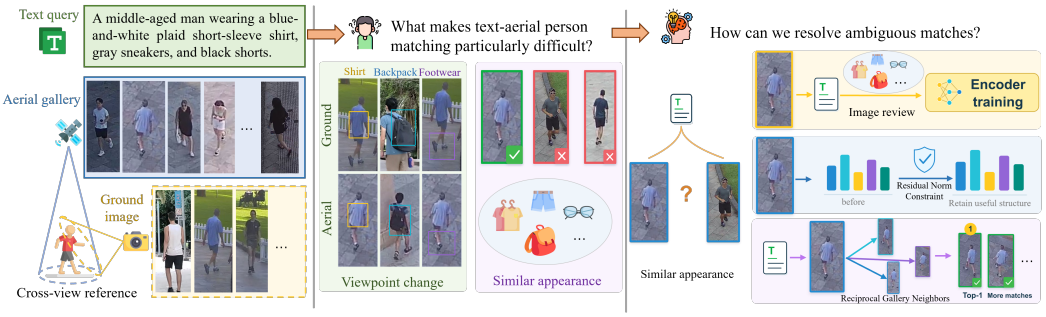
Text-aerial person retrieval aims to retrieve images of a target person from an aerial gallery given a natural language description. Variations in viewing angle, altitude, and occlusion can obscure the appearance cues described in the query, making visually similar individuals difficult to distinguish. Existing methods improve image-text alignment, but scoring each gallery image independently does not fully exploit context from other gallery images. In this paper, we propose Bounded Contextual Refinement (BCR), which improves aerial representations and retrieval rankings through bounded feature calibration and reciprocal gallery context. Bounded Feature Calibration (BFC) learns a norm-constrained residual transformation by contrasting the highest-scoring positive and negative gallery images for each text query. The constraint limits the magnitude of each feature correction while allowing the representation to adapt. Reciprocal Contextual Ranking (RCR) addresses two distinct retrieval decisions. It constructs candidate context vectors from reciprocal neighborhoods and learns bounded score corrections for candidate selection among the initial top five candidates. It then combines matching scores with neighborhood support to rank the remaining candidates within fixed intervals. To strengthen supervision for aerial appearance, Auxiliary Description Supervision (ADS) supplements training of the CLIP encoder with descriptions generated from aerial training images and reviewed against them by the same frozen vision-language model. Experiments on AERI-PEDES and TBAPR show improvements in Rank-1 accuracy and mAP over the CFAN baseline. The code will be made publicly available.

1 INTRODUCTION

Text-based person retrieval aims to retrieve images of a target person given a natural language description, without requiring a reference image (Li et al., 2017). Extending this task to aerial imagery enables person search across larger areas and varied viewpoints (Wang et al., 2026). As illustrated in Figure 1(a), ground-view images provide additional visual supervision during training, while retrieval uses a text query and an aerial gallery. Aerial observations often provide incomplete visual evidence. Greater flight altitudes reduce visible detail, oblique views change apparent body proportions, and occlusion can hide clothing or carried objects mentioned in a query (Deng et al., 2026). Figure 1(b) illustrates the resulting viewpoint differences and ambiguity between similar-looking individuals. Different identities may share the few attributes that remain visible, while separate images of the same person may reveal complementary attributes. Accurate retrieval therefore requires distinguishing visually similar individuals and combining evidence across gallery images.

Recent methods improve text-based person retrieval by adapting pretrained vision-language representations. CLIP provides a shared embedding space for images and text through large-scale contrastive pretraining (Radford et al., 2021). IRRA builds on this representation to learn fine-grained visual-textual relations (Jiang & Ye, 2023). Other approaches address noisy image-text correspondences or adapt the representation to domain differences (Qin et al., 2024, Liu et al., 2025). Attribute supervision strengthens the association between descriptions and visible appearance cues (Yang et al., 2023). Generated descriptions and diverse annotation patterns further enrich the language supervision available for retrieval (Tan et al., 2024, Jiang et al., 2025). These advances improve

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.



(a) Task Setup (b) Retrieval Challenges (c) Method Motivation

Figure 1. Text-aerial person retrieval and motivation for BCR. (a) Task setup with ground-view supervision during training. (b) Viewpoint variation and similar appearance make identities difficult to distinguish. (c) ADS, BFC, and RCR provide auxiliary text supervision, improve feature discrimination, and refine retrieval rankings. Rank-k is the percentage of queries with at least one correct match among the k highest-ranked gallery images.

image-text alignment, but incomplete aerial observations may still yield similar matching scores for different individuals with similar appearances.

Neighbor-aware correction uses local relationships to improve person retrieval (Wang et al., 2025b). Reciprocal neighborhoods retain image pairs that belong to each other's nearest-neighbor sets. However, visual proximity does not ensure a shared identity, so a neighbor with a different identity can increase the score of an incorrect candidate. Gallery context should therefore reflect the attributes specified by the query as well as visual similarity between images. Candidate selection requires distinguishing candidates with similar scores, while ranking the remaining candidates can benefit from query-specific support across gallery images. This distinction motivates separate, constrained uses of gallery context for the two decisions. Caption bootstrapping has shown the value of filtering generated descriptions before using them for vision-language learning (Li et al., 2022). For aerial retrieval, this review aims to exclude attributes unsupported by the source image.

In this paper, we propose Bounded Contextual Refinement (BCR) for text-aerial person retrieval, as illustrated in Figure 1(c). Bounded Feature Calibration (BFC) learns a norm-constrained residual using positive and negative gallery images for each text query, improving identity discrimination while limiting feature changes. The calibrated features support image-text matching and reciprocal gallery neighborhood construction in the same feature space. Reciprocal Contextual Ranking (RCR) uses these neighborhoods to select the top-ranked candidate from the initial top five through bounded score corrections. It then combines matching scores with neighborhood support to order the remaining images within fixed rank intervals, preserving the initial top-five and top-ten candidate sets. Auxiliary Description Supervision (ADS) supplements the released annotations with descriptions generated from aerial training images and reviewed against the same images by a frozen vision-language model. These descriptions are used only during training and require no language generation during retrieval. Our contributions are summarized below.

- We propose Bounded Feature Calibration (BFC) to learn norm-constrained residuals using positive and negative gallery images for each text query. The calibrated features support both image-text matching and reciprocal neighborhood construction.
- We propose Reciprocal Contextual Ranking (RCR) to use gallery context for candidate selection with bounded score corrections and ranking within fixed intervals. It refines the positions of relevant images while preserving the initial top-five and top-ten candidate sets.
- We propose Auxiliary Description Supervision (ADS) to adapt the CLIP encoder with aerial descriptions reviewed against their source images. These descriptions complement the released annotations and strengthen supervision for visible appearance cues.
- Extensive experiments on AERI-PEDES and TBAPR demonstrate the effectiveness of BCR, while ablation studies examine the contributions of individual components.

2 Related Work

Text-based and aerial-ground person retrieval. Text-based person retrieval was introduced with the CUHK-PEDES benchmark (Li et al., 2017). Early methods learned discriminative cross-modal projections for image-text matching (Zhang & Lu, 2018). Pose-guided attention and part-level alignment model fine-grained correspondences (Jing et al., 2020, Ding et al., 2021). Weak supervision and pseudo-text pretraining reduce annotation requirements (Zhao et al., 2021, Shao et al., 2023). CLIP-based methods adapt pretrained vision-language representations for text-based person retrieval (Radford et al., 2021, Cao et al., 2024). Detailed descriptions and stronger visual features further improve retrieval (Zuo et al., 2024, Shen et al., 2025). Other approaches address noisy correspondences and calibrate fine-grained alignment (Qin et al., 2024, Li et al., 2026). Relation-aware learning also models cross-modal associations (Bai et al., 2023). Interactive retrieval uses dialogue and user feedback (Bai et al., 2025b, Qin et al., 2025). Aerial-ground person re-identification addresses person matching across aerial and ground camera views (Zhang et al., 2024, Wang et al., 2025a). Recent work develops view-invariant representations and aerial-ground video benchmarks (Khalid et al., 2025, Nguyen et al., 2025). Text-aerial and text-based aerial-ground retrieval further combine these viewpoint challenges with language queries (Deng et al., 2026, Zhou et al., 2026).

Vision-language adaptation and gallery context. Parameter-efficient adapters specialize pretrained models with small additional modules (Houlsby et al., 2019). Feature adapters and support-set feature caches adapt vision-language representations to downstream tasks (Gao et al., 2024, Zhang et al., 2022). Domain-aware adapters address distribution differences in person retrieval (Liu et al., 2025). Prompt learning adapts textual context through learned or image-conditioned prompts (Zhou et al., 2022a,b). Multimodal prompt learning adapts both visual and textual representations (Khattak et al., 2023). Caption bootstrapping and generative vision-language models provide complementary language supervision (Li et al., 2022, 2023). Neighbor-aware correction further incorporates relationships between gallery images into image-text matching (Wang et al., 2025b). We bound the magnitude of aerial-feature corrections and use reciprocal gallery context for candidate selection and ranking within fixed intervals.

3 Method

3.1 Overview

Figure 2 illustrates BCR, which comprises ADS, BFC, and RCR. The CLIP model comprises a text encoder and a shared image encoder for aerial and ground-view images. The two CLIP encoder blocks depict the same model during adaptation and subsequent feature extraction. ADS fine-tunes the encoder with the CFAN objective and auxiliary supervision from descriptions reviewed by the generation model (Deng et al., 2026). BFC then calibrates aerial features, which RCR uses for candidate selection and interval-constrained ordering. The encoder, calibration module, and contextual scorers are trained sequentially, with previously trained modules frozen during subsequent stages.

Given a query q and an aerial gallery $\mathcal{G} = \{x_j\}_{j=1}^{N}$, the text and image branches f_T and f_A produce unit-normalized features $\mathbf{t} = f_T(q)/\left\|f_T(q)\right\|_2$ and $\mathbf{a}_j = f_A(x_j)/\left\|f_A(x_j)\right\|_2$ in $\mathbb{R}^D$. Calibration maps $\mathbf{a}_j$ to $\mathbf{b}_j = C_{\mathbf{W}}(\mathbf{a}_j)$, keeping text features fixed. Queries are evaluated independently against the same gallery. In Figure 2, token sequences $\mathbf{A}$ and $\mathbf{T}$ yield global-feature matrices $\mathbf{A}^c \in \mathbb{R}^{N\times D}$ and $\mathbf{T}^c \in \mathbb{R}^{Q\times D}$, where Q is the number of queries, D is the embedding dimension, and the superscript c identifies global feature representations. Row normalization gives $\overline{\mathbf{A}}^c$ and $\overline{\mathbf{T}}^c$, and calibrated rows form $\mathbf{B}^c$. The image and text features give the similarity matrix $\mathbf{S} = \overline{\mathbf{T}}^c(\mathbf{B}^c)^\top$.

3.2 Bounded Feature Calibration

We propose bounded feature calibration to improve discrimination between visually similar individuals. A shared linear map predicts an additive residual for each aerial feature, with its norm bounded relative to the input. An identity connection adds the input feature to the bounded residual before normalization. Text features remain fixed during calibration.

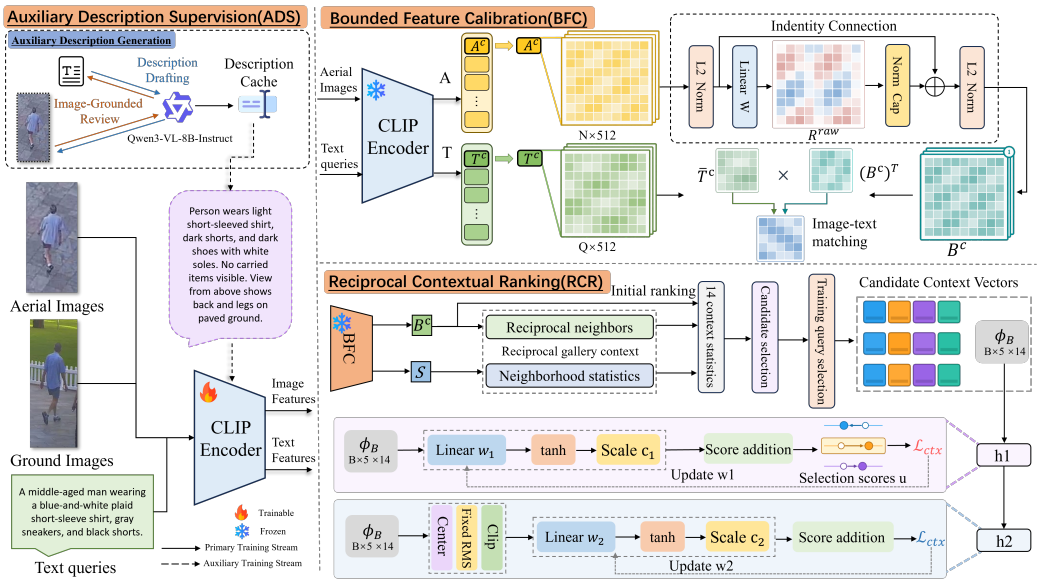


Figure 2. Overview of BCR. ADS adapts the CLIP encoder with auxiliary descriptions, BFC calibrates aerial features, and RCR uses reciprocal gallery context for candidate selection and ranking. Flames and snowflakes denote trainable and frozen modules, respectively.

Given a zero-initialized matrix $\mathbf{W} \in \mathbb{R}^{D\times D}$, we define the residual and calibrated feature as:

$$\mathbf{r}_{\mathbf{W}}(\mathbf{a}) = \mathbf{W}\mathbf{a}\min\left(1, \frac{\rho\,\|\mathbf{a}\|_2}{\max(\|\mathbf{W}\mathbf{a}\|_2, \epsilon)}\right), \qquad C_{\mathbf{W}}(\mathbf{a}) = \frac{\mathbf{a}+\mathbf{r}_{\mathbf{W}}(\mathbf{a})}{\|\mathbf{a}+\mathbf{r}_{\mathbf{W}}(\mathbf{a})\|_2}, \tag{1}$$

where ρ bounds the relative residual norm and ϵ ensures numerical stability. At initialization, the zero residual leaves each unit-normalized input unchanged. The residual satisfies $\|\mathbf{r}_{\mathbf{W}}(\mathbf{a})\|_2 \le \rho\,\|\mathbf{a}\|_2$. The norm cap rescales residuals that exceed the bound while preserving their direction. In Figure 2, $\mathbf{R}^{\text{raw}} = \overline{\mathbf{A}}^{c}\mathbf{W}^{\top}$ stacks the unbounded residuals. Final normalization places aerial and text features on the same unit sphere for both image-text matching and gallery similarity computation.

For each training query i, let p_i and n_i index its highest-scoring positive and negative gallery images. The calibration objective encourages a higher score for the selected positive image:

$$\mathcal{L}_{\text{cal}} = \frac{1}{B}\sum_{i=1}^{B} \text{softplus}\big(\mathbf{t}_i^{\top} C_{\mathbf{W}}(\mathbf{a}_{n_i}) - \mathbf{t}_i^{\top} C_{\mathbf{W}}(\mathbf{a}_{p_i})\big), \tag{2}$$

where B is the batch size and $\text{softplus}(x) = \log(1+e^x)$. We select these images at the start of each epoch and recompute their scores at each optimization step. Only $\mathbf{W}$ is trained, with the CLIP encoder frozen. The shared transformation supports both image-text matching and the construction of reciprocal neighborhoods among gallery images.

3.3 Reciprocal Contextual Ranking

RCR uses relationships among aerial gallery images for candidate selection and ranking within fixed intervals. Let s_j denote the initial image-text matching score of gallery candidate j, and let π be the ranking obtained by sorting these scores in descending order.

Reciprocal gallery context. We construct a gallery graph from calibrated aerial features, retaining connections between mutual nearest neighbors. Let $K_k(j)$ contain the k images most similar to image j, excluding j itself. The neighborhood and its query-dependent mean score are:

$$\mathcal{N}(j) = \{j\} \cup \{l \mid l \in K_k(j),\ j \in K_k(l)\}, \qquad \mu_j = \frac{1}{|\mathcal{N}(j)|}\sum_{l\in\mathcal{N}(j)} s_l, \tag{3}$$

where $k = 4$. Including the candidate itself ensures a nonempty neighborhood. The reciprocal graph remains fixed across queries. The neighborhood mean aggregates query-specific support from visually similar images, with higher matching scores increasing this mean.

Contextual candidate selection. For each candidate in the initial top $L = 64$ results, we construct a context vector $\boldsymbol{\phi}_j \in \mathbb{R}^F$ with $F = 14$. It combines score differences, neighborhood support, gallery similarity, initial rank, and neighborhood overlap with top-ranked candidates. Candidate selection is restricted to the set $\mathcal{S} = \{\pi_1, \ldots, \pi_M\}$ containing the initial top $M = 5$ candidates.

For a candidate $j = \pi_r$ at initial rank $r \le L$, let $n_j = |\mathcal{N}(j)|$ and let m_j and σ_j denote the maximum and population standard deviation of its neighborhood query scores. The mean μ_j is defined in Eq. 3. We define the gallery similarity statistics and neighborhood overlap as:

$$g_j^{\text{mean}} = \frac{1}{n_j} \sum_{l \in \mathcal{N}(j)} \mathbf{b}_j^\top \mathbf{b}_l, \quad g_j^{\min} = \min_{l \in \mathcal{N}(j)} \mathbf{b}_j^\top \mathbf{b}_l, \quad o_j^{(v)} = \frac{|\mathcal{N}(j) \cap \{\pi_1, \ldots, \pi_v\}|}{n_j}, \tag{4}$$

where v is the rank depth used to measure neighborhood overlap. Define the score statistics $\mathbf{d}_j = (s_j - s_{\pi_1}, \mu_j - s_j, m_j - s_j, \sigma_j)^\top$, gallery statistics $\mathbf{g}_j = (1 - g_j^{\text{mean}}, 1 - g_j^{\min})^\top$, and score gaps at selected ranks $\boldsymbol{\Delta}_j = (s_j - s_{\pi_{r+1}}, s_{\pi_1} - s_{\pi_2}, s_{\pi_1} - s_{\pi_M}, s_{\pi_1} - s_{\pi_L})^\top$. The context vector is:

$$\boldsymbol{\phi}_j = \text{clip}\big([\mathbf{d}_j^\top/\gamma_s, n_j/5, \mathbf{g}_j^\top/\gamma_g, \log r/\log L, \boldsymbol{\Delta}_j^\top/\gamma_s, o_j^{(8)}, o_j^{(L)}]^\top, -\kappa, \kappa\big), \tag{5}$$

where $\gamma_s = 0.05$ and $\gamma_g = 0.1$ scale query scores and gallery similarities, respectively, and $\text{clip}(\cdot, -\kappa, \kappa)$ clips each coordinate to $[-\kappa, \kappa]$ with $\kappa = 10$. The next-candidate gap at rank L uses rank $L + 1$, and neighborhoods may include images outside the initial context set.

The first scorer evaluates each candidate directly. For the second, we center each context within the selection set as $\mathbf{z}_j = \boldsymbol{\phi}_j - M^{-1} \sum_{l \in \mathcal{S}} \boldsymbol{\phi}_l$. Writing $\boldsymbol{\Phi} = (\boldsymbol{\phi}_l)_{l \in \mathcal{S}}$, the two score corrections are:

$$h_1(\boldsymbol{\phi}_j) = c_1 \tanh(\mathbf{w}_1^\top \boldsymbol{\phi}_j), \quad h_2(\boldsymbol{\Phi})_j = c_2 \tanh\big(\mathbf{w}_2^\top \text{clip}(\mathbf{z}_j \oslash \mathbf{v}, -\kappa, \kappa)\big), \tag{6}$$

where c_1 and c_2 bound the corrections, $\oslash$ denotes elementwise division, and $\mathbf{v}$ contains fixed root-mean-square scales estimated from all eligible centered training contexts. Each scale has a minimum value of 10^{-4}, and κ bounds the normalized coordinates. Both scorers use the initial contexts and zero-initialized weights $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^F$. Candidate selection uses the corrected matching score:

$$u_j = s_j + h_1(\boldsymbol{\phi}_j) + h_2(\boldsymbol{\Phi})_j, \qquad j^\star = \arg\max_{j \in \mathcal{S}} u_j. \tag{7}$$

We place the selected image first and preserve the relative order of the remaining candidates. Both corrections are computed before any candidate positions change.

Scorer training uses queries with both positive and negative images in the selection set. Let $\mathcal{P}_i$ and $\mathcal{N}_i$ denote the positive and negative candidate sets for query i, with $u_i^+ = \max_{j \in \mathcal{P}_i} u_{ij}$ and $u_i^- = \max_{j \in \mathcal{N}_i} u_{ij}$. The objective combines ranking penalties with correction regularization:

$$\mathcal{L}_{\text{ctx}} = \frac{1}{B} \sum_{i=1}^{B} \left(\ell_i^{\text{rank}} + \lambda_{\text{pos}} \ell_i^{\text{pos}} + \lambda_{\text{reg}} \ell_i^{\text{reg}}\right), \tag{8}$$

where $\ell_i^{\text{rank}} = \sigma((u_i^- - u_i^+)/\tau_1)$ compares the highest-scoring negative and positive candidates. The penalty averaged over positive candidates is $\ell_i^{\text{pos}} = |\mathcal{P}_i|^{-1} \sum_{j \in \mathcal{P}_i} \sigma((u_i^- - u_{ij})/\tau_2)$, and the regularizer is $\ell_i^{\text{reg}} = M^{-1} \sum_{j \in \mathcal{S}_i} (\delta_{ij}/c)^2$. Here σ is the sigmoid function, τ_1 and τ_2 are temperature parameters for the sigmoid penalties, and λ_{pos} and λ_{reg} weight the auxiliary terms. The scorer being optimized produces correction δ_{ij} with bound c. We train h_1 with $u = s + h_1$, then freeze it and fit h_2 with $u = s + h_1 + h_2$, regularizing only the second correction. For this stage, three disjoint groups of training identities define subgalleries. We reconstruct their graphs and contexts and fit one shared second scorer. Only $\mathbf{w}_2$ is updated during the second stage.

Interval-constrained ordering. After candidate selection, we combine individual and neighborhood evidence as $e_j = (1 - \alpha)s_j + \alpha\mu_j$, where α controls the neighborhood contribution. We fix the first result and sort the remaining candidates independently within positions $[2, 5]$, $[6, 10]$, and $[11, N]$ using e_j. Only candidate selection uses the learned corrections. Ties retain the order immediately before interval sorting.

Candidate selection is restricted to the initial top five, and subsequent sorting keeps candidates within their assigned intervals. RCR therefore preserves the initial top-five and top-ten candidate sets, leaving both Rank-5 and Rank-10 unchanged.

3.4 Auxiliary Description Supervision

ADS enriches aerial image-text training using a frozen Qwen3-VL-8B-Instruct model, which generates a description from an aerial training image and reviews it against the same image (Bai et al., 2025a). During review, the model checks clothing, footwear, and carried objects and is prompted to remove attributes unsupported by the image. Each description is cached with its source image path and training identity. Generation and review use no ground-view images or released captions. The fixed cache supplies auxiliary pairs during encoder adaptation.

The original image-text annotations remain the primary training data. Every m updates, an auxiliary batch pairs generated descriptions with their aerial images. We use bidirectional similarity distribution matching for global alignment (Jiang & Ye, 2023). The auxiliary loss combines this term with the local alignment objective retained from CFAN:

$$\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{SDM}}(\widetilde{\mathbf{T}}, \mathbf{A}) + \lambda_{\text{loc}}\mathcal{L}_{\text{local}}(\widetilde{\mathbf{T}}, \mathbf{A}), \qquad \mathcal{L}^{(u)} = \mathcal{L}_{\text{base}} + \lambda_{\text{aux}}^{(u)}\mathcal{L}_{\text{aux}}, \tag{9}$$

where $\widetilde{\mathbf{T}}$ and $\mathbf{A}$ denote features of generated descriptions and aerial images, $\mathcal{L}_{\text{base}}$ is the primary objective, and λ_{loc} weights local alignment. At update u, $\lambda_{\text{aux}}^{(u)}$ equals λ_{aux} when an auxiliary batch is sampled and zero otherwise. The primary loss is evaluated at every update and combined with the auxiliary loss, when sampled, in a single optimizer step. The auxiliary objectives use the same pairs and identity labels to define positive matches within a batch. BFC and RCR are subsequently trained with the released descriptions. Retrieval requires only the text query and aerial gallery.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets. We evaluate BCR on two text-aerial person retrieval benchmarks using their released splits. AERI-PEDES combines aerial and ground-view imagery from multiple sources, with generated training descriptions and manually annotated test queries (Deng et al., 2026). The released training records used in our experiments contain 3,522 identities. Its test set contains 6,141 text queries and 5,525 aerial gallery images from 1,000 identities. TBAPR contains 1,180 training identities and 529 test identities (Wang et al., 2026). We evaluate its 3,200 test queries against 3,200 gallery entries. Each query is evaluated against the complete test gallery of its dataset.

Evaluation metrics. We report Rank-k ($k = 1, 5, 10$), mean average precision (mAP), and RSum. Additional analyses use mean inverse negative penalty (mINP) (Ye et al., 2022). Coverage-64 is defined in Section 4.5. Rank-k, mAP, and mINP are percentages. RSum is computed by adding the percentage values of the Rank-1, Rank-5, and Rank-10 accuracies.

4.2 Implementation Details

We use a CLIP encoder initialized from the released CFAN checkpoint, retaining its primary training objective (Deng et al., 2026). Images are resized to 384×128, text sequences contain at most 77 tokens, and the embedding dimension is 512. We adapt the encoder for five epochs and train BFC for 49 epochs using Adam with a batch size of 64 (Kingma & Ba, 2014). Every eight optimization steps, ADS samples an auxiliary batch of 16 distinct identities. The residual bound is $\rho = 0.1$. The two RCR scorers are trained sequentially for 20 epochs each using AdamW, a batch size of 256, and a learning rate of 10^{-3} (Loshchilov & Hutter, 2017). We train a separate model for each dataset on an NVIDIA RTX 5090 GPU.

4.3 Comparison with State-of-the-Art Methods

Table 1 compares BCR with existing methods on both benchmarks. Results for the comparison methods are reported in the CFAN study (Deng et al., 2026).

Performance on AERI-PEDES. BCR achieves the best results across all five reported metrics. Rank-1 and mAP reach 49.29% and 47.70%, respectively, compared with 47.16% and 44.79% for CFAN with ground-view image supervision. Rank-5 and Rank-10 increase to 66.11% and 74.17%, yielding an RSum of 189.58. These gains indicate higher Rank-1 accuracy and improved overall

Table 1. Comparison on AERI-PEDES and TBAPR. Superscripts P and G denote additional pre-training and ground-view supervision, respectively. Best results are in bold.

Method	AERI-PEDES					TBAPR				
	Rank-1	Rank-5	Rank-10	mAP	RSum	Rank-1	Rank-5	Rank-10	mAP	RSum
IRRA (Jiang & Ye, 2023)	35.14	53.23	63.19	33.42	151.57	39.63	58.72	67.69	35.31	166.04
APTM (Yang et al., 2023)	34.62	53.95	64.50	31.09	153.07	43.59	62.03	69.75	38.71	175.37
RDE (Qin et al., 2024)	38.56	58.26	67.89	37.16	164.71	37.31	54.06	60.75	32.17	152.12
CFAM (Zuo et al., 2024)	30.77	51.37	61.61	30.40	143.75	48.34	66.31	73.21	42.67	184.78
NAM (Tan et al., 2024)	42.47	61.72	69.99	40.22	174.17	46.56	63.13	70.13	40.92	179.82
VFE (Shen et al., 2025)	35.76	55.35	65.56	35.35	156.67	47.94	62.50	68.17	42.18	178.63
DM-Adapter (Liu et al., 2025)	33.42	53.17	62.79	32.41	149.37	37.81	58.34	66.56	33.28	162.71
LPNC (Wang et al., 2025b)	35.65	53.61	63.69	35.19	152.95	41.78	58.03	65.50	37.87	165.31
LPNC^{P} (Wang et al., 2025b)	43.79	61.49	70.40	42.22	175.68	45.41	62.31	69.94	42.17	177.66
AEA-FIRM (Wang et al., 2026)	37.94	56.66	65.71	34.89	160.31	44.75	62.38	69.13	36.28	176.26
AEA-FIRM^{P} (Wang et al., 2026)	44.42	61.96	71.03	41.12	177.41	48.15	63.87	71.21	42.01	183.23
HAM (Jiang et al., 2025)	44.58	63.52	72.67	42.45	180.77	47.81	64.96	72.53	41.86	185.30
CFAN (Deng et al., 2026)	45.06	64.53	73.21	43.27	182.80	49.15	65.88	**73.47**	42.89	188.50
CFAN^{G} (Deng et al., 2026)	47.16	65.66	73.83	44.79	186.65	49.47	66.50	73.06	43.96	189.03
BCR (ours)	**49.29**	**66.11**	**74.17**	**47.70**	**189.58**	**49.72**	**66.75**	72.78	**44.87**	**189.25**

ranking quality. The ablations examine how feature calibration and gallery context affect Rank-1 accuracy and the positions of relevant images throughout the gallery.

Performance on TBAPR. BCR achieves the best Rank-1, Rank-5, mAP, and RSum among the compared methods, reaching 49.72%, 66.75%, 44.87%, and 189.25, respectively. Compared with CFAN with ground-view image supervision, mAP increases from 43.96% to 44.87%. Rank-10 is 72.78%, below the best baseline value of 73.47%. Figure 4(a) summarizes these metric-wise rankings on both datasets. The results show improvements in Rank-1 accuracy and overall ranking quality, with the largest gains observed on AERI-PEDES.

Figure 3 shows four selected AERI-PEDES queries as ADS, BFC, and RCR are added. Each row tracks the same four gallery images, and AP is computed over the complete gallery. For Q1, the incorrect candidate moves from rank 2 to ranks 4, 6, and 9, while AP rises from 79.68% to 100.00%. For Q4, RCR moves a correct match to the first position. These examples illustrate how the components affect the top-ranked result and the positions of other relevant images.

4.4 Ablation Study

Tables 2 to 5 examine the components of BCR on AERI-PEDES. The shaded rows in Tables 2 and 3 report the complete BCR method and reproduce the results in Table 1. The remaining configurations and additional controlled comparisons examine individual components under fixed conditions. When an ablation changes the representations, we retrain the affected downstream modules. Ranking rules use identical features and scorer weights.

Effectiveness of bounded feature calibration. Table 2 shows that BFC increases Rank-1 from 47.76% to 48.18% before RCR, while mAP remains at 45.61%. In an additional controlled comparison with RCR and ADS enabled, BFC raises Rank-1 from 48.64% to 49.00%. The corresponding mAP is 47.26%, compared with 47.33% without calibration. These results indicate that BFC primarily improves Rank-1 accuracy in this comparison, consistent with its objective of separating the highest-scoring positive and negative candidates. Rank-10 decreases from 74.34% to 74.03% with calibration. Optimizing the score difference between the highest-scoring positive and negative images does not constrain the relative positions of all other relevant images.

Effectiveness of reciprocal contextual ranking. Table 3 evaluates the two scorers and interval-constrained ordering, denoted by Order. With BFC and ADS fixed, candidate selection raises Rank-1 from 48.18% to 49.00%, while mAP changes from 45.61% to 45.71%. Interval ordering alone raises mAP to 47.16% without changing Rank-1. Combining both operations yields 47.26% mAP

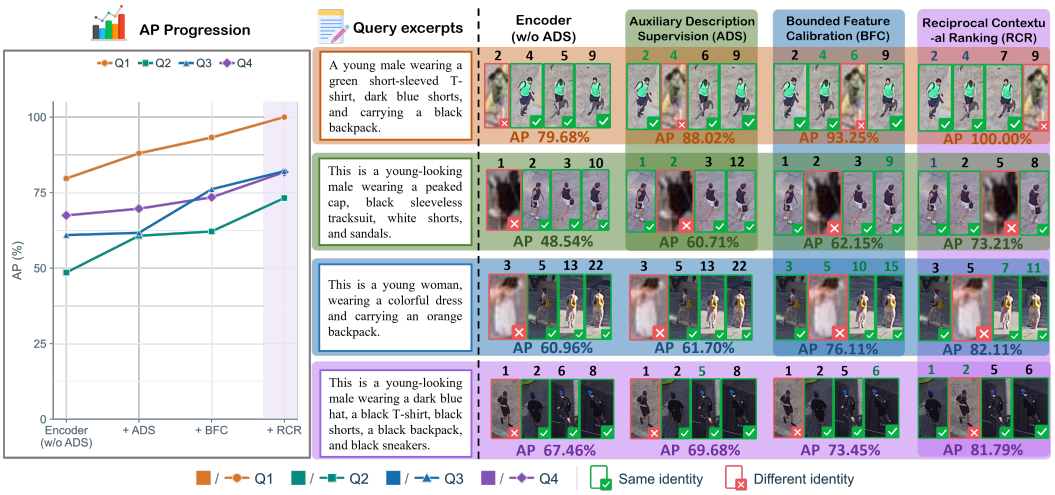


Figure 3. Retrieval results for four AERI-PEDES queries in the controlled component comparison. The left panel reports full-gallery AP after each component is added. The right panel tracks the same four gallery images at their actual ranks, which need not be consecutive. Green checks and red crosses indicate matching and nonmatching identities, respectively.

Table 2. Component analysis of BCR.

BFC	RCR	ADS	Rank-1	Rank-5	Rank-10	mAP	RSum
×	×	✓	47.76	65.89	**74.34**	45.61	187.98
✓	×	✓	48.18	65.79	74.03	45.61	188.00
×	✓	✓	48.64	65.89	**74.34**	47.33	188.86
✓	✓	×	48.64	65.66	74.04	47.05	188.34
✓	✓	✓	**49.29**	**66.11**	74.17	**47.70**	**189.58**

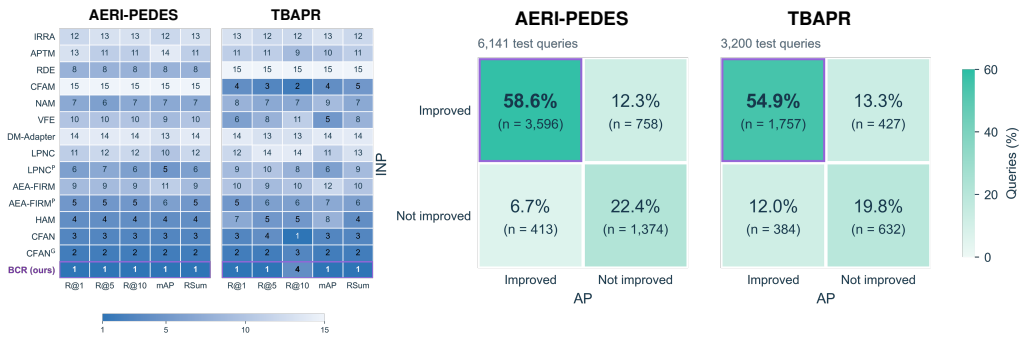
Table 3. Component analysis of RCR.

h_1	h_2	Order	Rank-1	Rank-5	Rank-10	mAP	RSum
×	×	×	48.18	65.79	74.03	45.61	188.00
✓	✓	×	49.00	65.79	74.03	45.71	188.81
×	×	✓	48.18	65.79	74.03	47.16	188.00
✓	×	✓	48.85	65.79	74.03	47.25	188.67
✓	✓	✓	**49.29**	**66.11**	**74.17**	**47.70**	**189.58**

in the controlled comparison. Adding h_2 to h_1 with interval ordering increases Rank-1 from 48.85% to 49.00%, with a smaller mAP change from 47.25% to 47.26%. Candidate selection accounts for the Rank-1 improvement, while interval ordering provides most of the mAP gain. Selection changes only the order of the initial five candidates. Interval ordering also adjusts positions below rank 10, allowing additional relevant images to contribute to the mAP improvement. Rank-5 and Rank-10 remain unchanged when the same initial scores are used because RCR preserves both candidate sets.

Effectiveness of the residual bound. Table 4 compares bounded and unbounded residual transformations with RCR and ADS retained. Bounding the residual improves Rank-1 from 48.53% to 49.00% and mAP from 46.68% to 47.26%. Rank-5 and Rank-10 also increase, producing an RSum of 188.81 instead of 187.64. The consistent gains across these metrics support constraining the magnitude of feature adaptation. The residual can improve identity discrimination while limiting changes to the pretrained image-text representation. Both variants use the same linear residual parameterization, so enforcing the norm constraint introduces no additional learned parameters. The benefit therefore comes from restricting feature changes within this parameterization.

Effectiveness of auxiliary description supervision. An additional controlled comparison shows that ADS raises Rank-1 from 48.64% to 49.00% and mAP from 47.05% to 47.26%. Table 5 further compares released and generated descriptions using a fixed total of 1,021 auxiliary records from either 256 or 1,021 training identities. For 256 identities, generated descriptions increase mAP from 47.08% to 47.26%. For 1,021 identities, mAP increases from 47.07% to 47.19%, and mINP rises from 35.03% to 35.24%. Generated descriptions improve all four reported metrics at either identity count. At 256 identities, mINP increases from 35.05% to 35.16%, indicating a small improvement in the position of the last relevant image. Using the same number of auxiliary records makes the two description sources directly comparable. With the record count fixed, broader identity coverage yields mixed changes across metrics.



(a) Metric-wise method ranking

(b) Joint query-level AP and INP changes

Figure 4. Benchmark rankings and query-level effects of RCR. (a) Metric-wise ranks of the 15 configurations in Table 1, where 1 is best. (b) Joint AP and INP changes under the settings of Table 6, with query percentages and counts. A query is counted as improved when the metric increases by more than 10^{-4} percentage points. The remaining queries have unchanged or decreased values.

Table 4. Effect of the residual norm constraint.

Residual	Rank-1	Rank-5	Rank-10	mAP	RSum
Unbounded	48.53	65.23	73.88	46.68	187.64
Bounded	**49.00**	**65.79**	**74.03**	**47.26**	**188.81**

Table 5. Effect of auxiliary description sources.

Identities	Source	Rank-1	mAP	mINP	RSum
256	Released	48.80	47.08	35.05	188.52
256	Generated	**49.00**	**47.26**	**35.16**	**188.81**
1,021	Released	48.75	47.07	35.03	188.76
1,021	Generated	**48.80**	**47.19**	**35.24**	**189.06**

Table 6. Effect of RCR on retrieval.

RCR	Rank-1	mAP	mINP	Coverage-64
AERI-PEDES				
×	48.67	46.10	32.92	79.18
✓	**49.29**	**47.70**	**35.86**	**80.41**
Δ	+0.62	+1.60	+2.94	+1.22
TBAPR				
×	49.56	43.79	25.34	70.66
✓	**49.72**	**44.87**	**27.22**	**71.32**
Δ	+0.16	+1.08	+1.87	+0.66

4.5 Retrieval Performance Analysis

Table 6 evaluates RCR with fixed encoder, calibration, and scoring settings. RCR raises mAP from 46.10% to 47.70% on AERI-PEDES and from 43.79% to 44.87% on TBAPR. The corresponding mINP values increase from 32.92% to 35.86% and from 25.34% to 27.22%. Coverage-64 also rises from 79.18% to 80.41% on AERI-PEDES and from 70.66% to 71.32% on TBAPR. INP divides the number of relevant images by the rank of the last relevant image. For each query, Coverage-64 is the fraction of its relevant gallery images retrieved within the top 64 positions. We report the mean over queries. The gains in both measures indicate earlier retrieval of additional relevant images. Higher coverage reflects more relevant images occurring before the cutoff within the same gallery. Rank 64 lies within the final sorting interval, so Coverage-64 can increase while the top-five and top-ten candidate sets remain unchanged. All differences Δ use unrounded results.

Figure 4(b) shows that AP and INP both improve for 3,596 of 6,141 AERI-PEDES queries (58.6%) and 1,757 of 3,200 TBAPR queries (54.9%). AP alone improves for a further 6.7% and 12.0%, respectively. Both metrics therefore improve for a majority of queries on each benchmark. Other queries improve on only one metric or have unchanged or decreased values for both. AP can increase when earlier matches move forward even if the last relevant rank, and hence INP, remains unchanged.

5 Conclusion

We present Bounded Contextual Refinement for text-aerial person retrieval. BFC learns norm-constrained residuals to adapt aerial features, and RCR uses reciprocal gallery evidence for candidate selection and interval-constrained ordering. ADS complements the released annotations with descriptions reviewed against aerial training images. Experiments on AERI-PEDES and TBAPR show higher Rank-1 accuracy and mAP than the CFAN baseline. Ablations examine the effects of feature calibration, candidate selection, and interval ordering.

AI USE STATEMENT

Generative AI tools assisted the development and review of the conceptual framework, mathematical formulation, experimental design, method implementation, and interpretation of results. Qwen3-VL-8B-Instruct generated auxiliary descriptions from training images, as described in Section 3.4. Generative AI tools also supported literature searches, manuscript drafting and editing, figure design, and reference formatting. We checked AI-assisted formulations and code against the retrieval pipeline and verified table values against saved experimental outputs. We take responsibility for the final manuscript, including all AI-assisted text, results, and research artifacts.

ETHICS STATEMENT

The experiments use existing research benchmarks under their access terms. Person retrieval can be misused for non-consensual surveillance, and a retrieved match should not be treated as a verified identity. Our evaluation measures retrieval accuracy on these benchmarks. It does not establish deployment safety, consent in other settings, or demographic fairness.

REPRODUCIBILITY STATEMENT

Section 3 defines the modules, objectives, and training sequence. Section 4.2 summarizes the implementation used for both benchmarks. Section 4.4 identifies the active components and controlled conditions in each ablation. We retain the experimental configurations and raw evaluation outputs used to verify all reported metrics and component comparisons.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. RaSa: Relation and sensitivity aware representation learning for text-based person search. In Edith Elkind (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 555–563. International Joint Conferences on Artificial Intelligence Organization, 2023. doi: 10.24963/ijcai.2023/62.

Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, and Mang Ye. Chat-based person retrieval via dialogue-refined cross-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3952–3962, 2025b.

Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of CLIP for text-based person search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(1):465–473, 2024. doi: 10.1609/aaai.v38i1.27801.

Yifei Deng, Chenglong Li, Yuyang Zhang, Guyue Hu, and Jin Tang. Cross-modal fuzzy alignment network for text-aerial person retrieval and a large-scale benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38732–38741, 2026.

Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*, 2021.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019.

Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2787–2797, 2023.

Jiayu Jiang, Changxing Ding, Wentao Tan, Junhong Wang, Jin Tao, and Xiangmin Xu. Modeling thousands of human annotators for generalizable text-to-image person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 9220–9230, 2025.

Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. Pose-guided multi-granularity attention network for text-based person search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11189–11196, 2020. doi: 10.1609/aaai.v34i07.6777.

Wajahat Khalid, Bin Liu, Xulin Li, Muhammad Waqas, and Muhammad Sher Afgan. Bridging the sky and ground: Towards view-invariant feature learning for aerial-ground person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9749–9758, 2025.

Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19113–19122, 2023.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Dongsheng Li, Xinyuan Guo, Huijie Zhang, Pingting Hao, and Qiushi Xia. Quota-calibrated fine-grained alignment with context-aware marginals for text-based person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31103–31112, 2026.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5187–5196, 2017. doi: 10.1109/CVPR.2017.551.

Yating Liu, Zimo Liu, Xiangyuan Lan, Wenming Yang, Yaowei Li, and Qingmin Liao. DM-Adapter: Domain-aware mixture-of-adapters for text-based person retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):5703–5711, 2025. doi: 10.1609/aaai.v39i6.32608.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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 (CVPR)*, pp. 1241–1251, 2025.

Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27197–27206, 2024.

Yang Qin, Chao Chen, Zhihang Fu, Dezhong Peng, Xi Peng, and Peng Hu. Human-centered interactive learning via MLLMs for text-to-image person re-identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 14390–14399, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, and Jingdong Wang. Unified pre-training with pseudo texts for text-to-image person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11174–11184, 2023.

Wei Shen, Ming Fang, Yuxia Wang, Jiafeng Xiao, Diping Li, Huangqun Chen, Ling Xu, and Weifeng Zhang. Enhancing visual representation for text-based person searching. *Knowledge-Based Systems*, 309:112893, 2025.

Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. Harnessing the power of MLLMs for transferable text-to-image person ReID. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17127–17137, 2024.

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 Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 22119–22128, 2025a.

Xueping Wang, Hao Wu, Min Liu, and Yaonan Wang. Learnable prompts with neighbor-aware correction for text-based person search. *IEEE Transactions on Information Forensics and Security*, 20:8951–8961, 2025b. doi: 10.1109/tifs.2025.3598619.

Yihao Wang, Meng Yang, Rui Cao, and Guangwei Gao. AEA-FIRM: Adaptive elastic alignment with fine-grained representation mining for text-based aerial pedestrian retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(1):900–912, 2026. doi: 10.1109/tcsvt.2025.3586601.

Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4492–4501, 2023.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022. doi: 10.1109/tpami.2021.3054775.

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

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-Adapter: Training-free adaption of CLIP for few-shot classification. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 493–510, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19833-5.

Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Shizhen Zhao, Changxin Gao, Yuanjie Shao, Wei-Shi Zheng, and Nong Sang. Weakly supervised text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11395–11404, 2021.

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 (CVPR)*, 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. doi: 10.1007/s11263-022-01653-1.

Xinyu Zhou, Yu Wu, Jiayao Ma, Wenhao Wang, Min Cao, and Mang Ye. Text-based aerial-ground person retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(34):29035–29043, 2026. doi: 10.1609/aaai.v40i34.40140.

Jialong Zuo, Hanyu Zhou, Ying Nie, Feng Zhang, Tianyu Guo, Nong Sang, Yunhe Wang, and Changxin Gao. UFineBench: Towards text-based person retrieval with ultra-fine granularity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22010–22019, 2024.

APPENDIX

This appendix provides implementation details, structural properties, repeated training results, additional ranking analyses, computational cost measurements, and qualitative examples. Results from the checkpoint used for the main experiments are distinguished from the separate experiments with three training seeds.

Contents **Page**

A Implementation details 15
A.1 Data and evaluation protocol 15
A.2 Encoder adaptation and auxiliary descriptions 15
A.3 Bounded feature calibration 16
A.4 Reciprocal neighborhoods and context features 16
A.5 Scorer training and final ranking 17
B Structural properties 18
B.1 Bound on feature changes 18
B.2 Preservation of candidate sets 18
C Repeated training runs 19
C.1 Protocol and complete models 19
C.2 Component comparisons 20
D Additional ranking analyses 20
D.1 Distribution of changes across queries 20
D.2 Paired AP and INP distributions 22
D.3 Changes in Rank-1 retrieval accuracy 22
E Computational cost 22
E.1 Measurement protocol 22
E.2 Measured cost and scaling 23
F Qualitative analysis 25
F.1 Training images and auxiliary descriptions 25
F.2 Retrieval improvements 26
F.3 Failure cases 27

A Implementation details

A.1 Data and evaluation protocol

We use the released training and test partitions of AERI-PEDES and TBAPR, as in the main paper. Table A1 gives the number of training records, identities, test queries, and gallery instances. Each test query is evaluated against the complete gallery. The released training records may contain repeated image paths, so record counts should not be interpreted as counts of unique images. The two benchmarks share some source imagery and do not constitute a test of transfer to independent data sources.

Table A1. Dataset partitions used in the experiments. Training counts refer to released records, and gallery counts refer to evaluated instances.

Partition	AERI-PEDES	TBAPR
Training records	76,007	31,340
Training identities	3,522	1,180
Test queries	6,141	3,200
Gallery instances	5,525	3,200
Test identities	1,000	529

For a query with M matching gallery instances, let y_r indicate whether the instance at rank r has the same identity. Average precision and inverse negative penalty are computed over the full gallery as

$$\mathrm{AP} = \frac{1}{M}\sum_{r=1}^{N} y_r \frac{\sum_{u=1}^{r} y_u}{r}, \qquad \mathrm{INP} = \frac{M}{\max\{r \mid y_r = 1\}}. \tag{A.1}$$

All evaluated queries have at least one matching instance. Rank-k measures whether a match occurs within the first k positions. We average these quantities over queries and report percentages. RSum is the sum of Rank-1, Rank-5, and Rank-10. Coverage-k is the mean fraction of relevant instances retrieved within the top k positions, computed separately for each query. These metrics use identity labels only for evaluation.

The complete BCR results in the component tables match the main benchmark evaluation. Additional component comparisons use fixed evaluation conditions to isolate the changes under study. The accompanying experimental records retain the configuration associated with every reported result.

A.2 Encoder adaptation and auxiliary descriptions

We initialize each model from its corresponding pretrained CFAN checkpoint and continue training the full model for five epochs. Images are resized to 384×128, text sequences contain at most 77 tokens, and global embeddings have 512 dimensions. We retain the inherited local alignment modules, with four local queries for AERI-PEDES and twelve for TBAPR. The primary objective is CDA plus the FTA objective weighted by 0.5, following the CFAN implementation (Deng et al., 2026).

Every eighth global update includes an auxiliary batch with 16 distinct training identities. This batch uses the SDM objective and FTA weighted by 0.5. Its total weight is 1. The primary and auxiliary objectives are averaged separately within their respective batches. Their gradients are accumulated before a single optimizer update. The auxiliary counter continues across epoch boundaries. The five epochs contain 5,940 primary updates and 742 auxiliary batches for AERI-PEDES, and 2,450 primary updates and 306 auxiliary batches for TBAPR. The final incomplete primary batch is retained.

We use Adam with batch size 64, $(\beta_1, \beta_2) = (0.9, 0.999)$, numerical-stability parameter 10^{-3}, and weight decay 4×10^{-5}. The peak learning rates are 5×10^{-7} for the image and text backbones, 2.5×10^{-5} for the local-query parameters, and 2.5×10^{-6} for the remaining parameters. At the start of epoch e, each peak rate is multiplied by

$$f_e = 0.1 + 0.45\left[1 + \cos\left(\frac{\pi(e-1)}{4}\right)\right], \qquad e \in \{1, \ldots, 5\}. \tag{A.2}$$

The gradient norm is clipped to 5. Model parameters are stored in FP32. Backbone forward passes use BF16 autocasting, while normalization and numerically sensitive calculations use FP32.

Description generation. We generate auxiliary descriptions before training using a frozen Qwen3-VL-8B-Instruct model (Bai et al., 2025a). A first pass generates a draft from an aerial training image. A second pass reviews the draft against the same image. Ground-view images and released text annotations are excluded from both generation passes. The prompts focus on observable clothing, footwear, and carried objects. They ask the model to omit uncertain details, background information, and demographic attributes, and to use approximately 35 to 50 words when sufficient visual detail is available. This model-based review is not a substitute for human verification of every generated detail.

Generation uses batch size 2, BF16, scaled dot-product attention, deterministic decoding, and at most 160 new tokens per pass. The image processor uses a minimum of 131,072 pixels and a maximum of 262,144 pixels. The random seed is 116001. The exact model revision, generation prompts, and settings are included in the accompanying data files. Descriptions remain fixed throughout encoder training. The generation model is not used during retrieval.

Table A2. Auxiliary training data used by ADS. Repeated image paths are retained as distinct sampled records when present.

Quantity	AERI-PEDES	TBAPR
Sampled identities	256	86
Description records	1,021	328
Distinct aerial image paths	972	328
Auxiliary record exposures	11,872	4,896

Sampling is stratified by the source datasets represented in the released training partition. We sample 86, 85, and 85 identities from the three AERI-PEDES sources and 43 identities from each of the two TBAPR sources. Each identity contributes at most four records, with distinct aerial image paths preferred. Table A2 distinguishes the size of the fixed description cache from the number of record exposures during training. The primary training set is unchanged. Within each auxiliary batch, identities are distinct, while different cached records for the same identity may be sampled in successive batches.

A.3 Bounded feature calibration

After encoder adaptation, we freeze the encoder and train BFC for 49 epochs. The residual matrix $W \in \mathbb{R}^{512\times512}$ has no bias and is initialized to zero. The residual bound is $\rho = 0.1$, and the denominator in the residual rescaling uses a numerical floor of 10^{-12}. Positive and negative image indices are mined at the start of each epoch from the full training feature pool. For each text query, the selected positive has the highest current score among images with the same identity as the query, and the selected negative has the highest score among images with a different identity. These indices are fixed within the epoch, but scores are recomputed with the current calibration matrix at each update.

BFC uses Adam with batch size 64, the same β_1 and β_2 values, numerical-stability parameter, and weight decay as encoder adaptation, and the epoch schedule

$$\eta_e = \frac{2.5\times10^{-5}}{2}\left[1+\cos\left(\frac{\pi(e+5)}{55}\right)\right], \qquad e\in\{1,\dots,49\}. \tag{A.3}$$

The loss is the mean softplus of the negative score minus the positive score, with no additional temperature. The stored calibration pools contain 112,672 AERI-PEDES feature records and 31,340 TBAPR feature records. Training performs 58,212 and 24,010 updates, respectively. We preserve the query orders associated with the original training runs rather than replacing them with newly generated permutations.

A.4 Reciprocal neighborhoods and context features

The gallery graph is constructed from calibrated aerial image features. For each gallery instance, we select its four nearest neighbors by cosine similarity, excluding itself. An edge is retained only

if the relation is reciprocal. We then add the instance itself to its neighborhood. Neighborhoods therefore contain between one and five instances. Ties are resolved by the original gallery order. No test identity labels or query features are used to construct this graph.

Let π denote the stable descending order of the initial scores s, and let $j = \pi_r$ be the candidate at rank r. For its neighborhood $\mathcal{N}_j$, define $n_j = |\mathcal{N}_j|$. The quantities μ_j, m_j, and σ_j are the mean, maximum, and population standard deviation of the query scores within that neighborhood. The quantities g_j^{mean} and $g_j^{\min}$ are the mean and minimum cosine similarities between the calibrated feature of j and the features in its neighborhood. All these statistics include the candidate itself. Finally, $o_j^{(d)}$ is the fraction of its neighbors that occur within the initial top d ranks.

Table A3. The 14 context coordinates used by both RCR scorers. Each coordinate is clipped to $[-10, 10]$. Ranks are indexed from 1.

Coordinate	Value	Coordinate	Value
1	$(s_j - s_{\pi_1})/0.05$	8	$\log(r)/\log(64)$
2	$(\mu_j - s_j)/0.05$	9	$(s_{\pi_r} - s_{\pi_{r+1}})/0.05$
3	$(m_j - s_j)/0.05$	10	$(s_{\pi_1} - s_{\pi_2})/0.05$
4	$\sigma_j/0.05$	11	$(s_{\pi_1} - s_{\pi_5})/0.05$
5	$n_j/5$	12	$(s_{\pi_1} - s_{\pi_{64}})/0.05$
6	$(1 - g_j^{\text{mean}})/0.1$	13	$o_j^{(8)}$
7	$(1 - g_j^{\min})/0.1$	14	$o_j^{(64)}$

The context vector $\phi_j \in \mathbb{R}^{14}$ is computed from the initial ranking and is shared by both scorers. Table A3 specifies the coordinates for the initial top 64 candidates. The gap for the candidate at rank 64 uses the score at rank 65. The features are not recomputed after the first scorer changes a score. The graph and gallery cosine statistics depend only on the calibrated aerial features. Score statistics also depend on the text query.

A.5 Scorer training and final ranking

Each scorer contains 14 zero-initialized weights and no bias. The first produces a correction $\delta_{1j} = 0.05 \tanh(w_1^\top \phi_j)$. Before the second scorer is applied, each context coordinate is centered across the initial top five candidates of the same query. If $z_{ij\ell}$ is the centered coordinate for query i, candidate j, and coordinate ℓ, the fixed training scale is

$$v_\ell = \max\left\{\sqrt{\frac{1}{5Q_2}\sum_{i=1}^{Q_2}\sum_{j=1}^{5} z_{ij\ell}^2},\, 10^{-4}\right\}, \qquad \delta_{2j} = 0.02 \tanh\left(w_2^\top \operatorname{clip}(z_j/v, -10, 10)\right), \tag{A.4}$$

where Q_2 is the number of eligible training queries for the second scorer. The division is coordinatewise. The scales are computed once from its training data and retained at test time.

Only queries whose initial top five candidates contain both a matching and a nonmatching identity are eligible for scorer training. The objective combines a sigmoid penalty comparing the highest-scoring negative and positive candidates, a mean sigmoid penalty comparing that negative with every positive candidate, and a quadratic penalty on the correction being optimized. The two sigmoid temperatures are 0.02 and 0.03. The latter two terms have weights 0.1 and 0.005. The mean positive-candidate loss includes the highest-scoring positive and uses all positives in the eligible top-five set. It is not a full-gallery AP objective.

We train the first scorer for 20 epochs, freeze it, and then train the second for 20 epochs. Both use AdamW with batch size 256, learning rate 10^{-3}, weight decay 10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.999)$, and numerical-stability parameter 10^{-8} (Loshchilov & Hutter, 2017). The learning rate is constant. The second scorer is trained using three disjoint groups of training identities. Each query is paired only with the subgallery containing its identity, and all eligible queries are visited once per epoch. These groups alter the training gallery composition. They are not validation identities unseen by the encoder.

For the checkpoint used for the main results, the first scorer uses 29,484 eligible AERI-PEDES queries and 7,906 eligible TBAPR queries, corresponding to 2,320 and 620 updates. The second uses 27,900 and 7,352 eligible queries, corresponding to 2,180 and 580 updates. Eligibility depends on the learned representation, so the counts can differ in the repeated training runs.

For each query, the candidate with the largest $s_j + \delta_{1j} + \delta_{2j}$ among the initial top five is moved to rank 1. The other candidates initially retain their relative order. We then keep the top-ranked candidate fixed and sort the remaining candidates independently within ranks 2 to 5, 6 to 10, and 11 to N. Each interval uses $e_j = 0.75s_j + 0.25\mu_j$, which excludes the learned corrections. All sorting is stable. The output is a complete permutation of the original gallery instances.

B STRUCTURAL PROPERTIES

We establish two properties that follow directly from the construction. They describe the magnitude of feature changes and the preservation of selected candidate sets. Neither property guarantees that retrieval accuracy increases.

B.1 BOUND ON FEATURE CHANGES

Proposition B.1. Let a be a unit-norm feature vector, let r satisfy $\|r\|_2 \leq \rho < 1$, and let $b = (a + r)/\|a + r\|_2$. Then b is well defined and $\|b - a\|_2 \leq 2\rho$.

Proof. The reverse triangle inequality gives $\|a + r\|_2 \geq 1 - \rho > 0$. Moreover,

$$\begin{aligned} \|b - a\|_2 &\leq \|b - (a + r)\|_2 + \|r\|_2 \\ &= |1 - \|a + r\|_2| + \|r\|_2 \\ &\leq 2\|r\|_2 \leq 2\rho. \end{aligned} \tag{A.5}$$

The BFC rescaling enforces the assumed residual bound, including when the unscaled residual is zero. This establishes the claim. □

For a unit text feature t, the corresponding score difference satisfies $|t^\top b - t^\top a| \leq 2\rho$. This is a bound on a score change. Small score changes can still reverse the order of closely ranked candidates, so the bound does not imply preservation of identity predictions or retrieval ranks.

B.2 PRESERVATION OF CANDIDATE SETS

Proposition B.2. Let π be the ranking immediately before RCR. The final top-five and top-ten candidate sets are identical to the corresponding sets in π.

Proof. The candidate selection stage chooses a candidate from the initial top five. Moving that candidate to rank 1 permutes only those five positions. Sorting ranks 2 to 5 then permutes the other four candidates within the same set. The candidates initially at ranks 6 to 10 are permuted only within that interval, and candidates below rank 10 cannot enter it. The union of the first two sets is therefore also unchanged. □

Consequently, Rank-5, Rank-10, Coverage-5, and Coverage-10 are unchanged by RCR relative to its own input ranking. Rank-1, AP, and INP can increase or decrease because the order within these sets and below rank 10 can change. Feature calibration and initial score computation precede RCR and can change the input sets. The proposition therefore does not assert that the full method preserves the top-five or top-ten sets of the original encoder.

Figure A1 illustrates the distinction between forming reciprocal neighborhoods and preserving ranking intervals. The gallery graph determines which image scores contribute to each neighborhood statistic. The interval boundaries instead constrain where candidates may move after the first candidate is selected. Reciprocal neighbors can have different identities, so the graph does not itself establish that a candidate is a correct match.

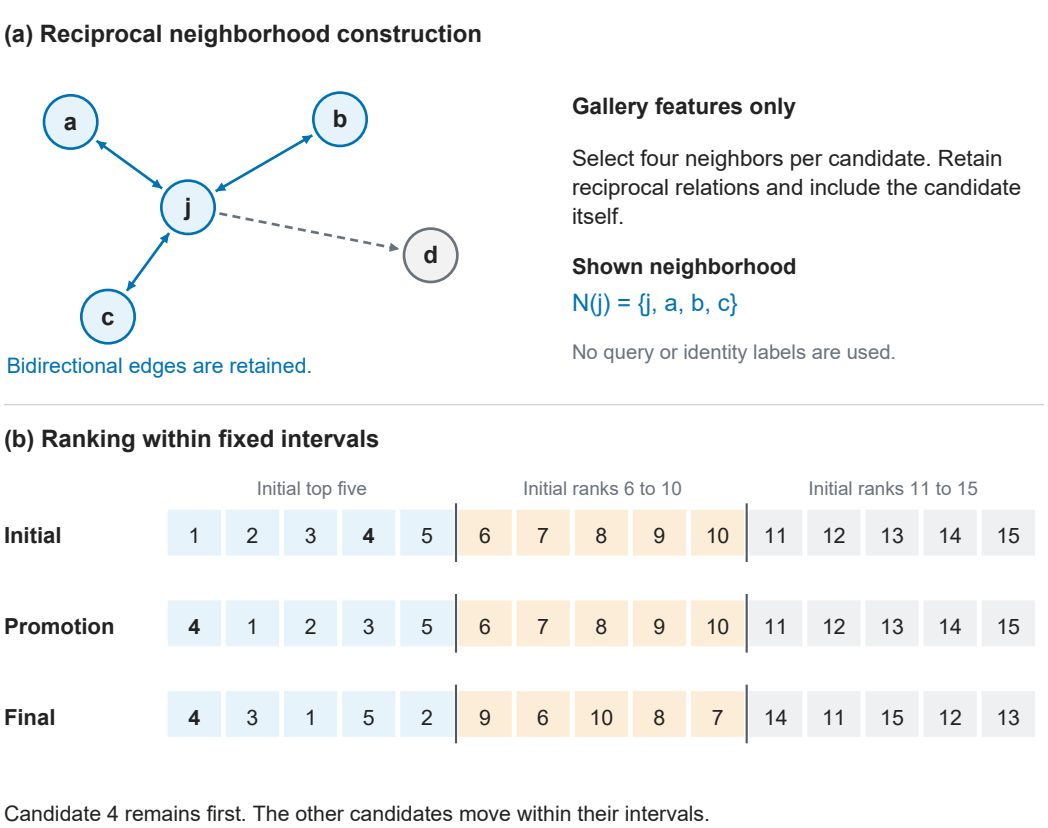


Figure A1. Schematic illustration of reciprocal context and interval ordering. Panel (a) shows selected gallery relations. Reciprocal edges with a, b, and c are retained, while the one-way relation with d is excluded. The candidate j is included in its own neighborhood. In (b), the numbers denote candidate indices in a constructed gallery of 15 instances. Candidate 4 is promoted from the initial top five and remains first during subsequent ordering. Colors indicate the initial rank intervals. This constructed example illustrates the ranking rule and does not report measured retrieval results.

C Repeated Training Runs

C.1 Protocol and Complete Models

We repeat the training procedure with seeds 1901, 1902, and 1903. Each run starts from the same dataset-specific pretrained CFAN checkpoint and repeats encoder adaptation, BFC fitting, and RCR training. The description cache and identity partitions remain fixed. These runs measure variability in the training stages following initialization from the fixed CFAN checkpoint. They do not include variability from training CFAN from scratch. The checkpoint used for the main results remains fixed and is not replaced by the best-performing repeated run.

Table A4 reports the mean and sample standard deviation with denominator $n - 1$. The accompanying data contain each seed and all six metrics, including RSum. The small number of repetitions does not support broad claims about statistical significance. In particular, test queries are not treated as independent repetitions of training.

In the controlled comparison, ADS increases AERI-PEDES mAP by a mean of 0.142 percentage points relative to the corresponding runs without ADS, with a sample standard deviation of 0.028 for the paired difference. This mAP difference is positive in all three pairs. Rank-1 does not improve in every pair, and mean Rank-5 decreases by 0.195 points. These results show that a change that improves mAP need not improve every rank metric.

Table A4. Repeated training results reported as mean and sample standard deviation over three seeds. Rows marked *Controlled* report fixed conditions for component analysis. Rows marked *Main setting* follow the main benchmark evaluation. All values are percentages.

Setting	Rank-1	Rank-5	Rank-10	mAP	mINP
AERI-PEDES					
Controlled	48.82 ± 0.02	65.73 ± 0.10	74.17 ± 0.06	47.22 ± 0.01	35.15 ± 0.06
Main setting	49.19 ± 0.06	65.90 ± 0.08	74.34 ± 0.16	47.69 ± 0.02	35.80 ± 0.07
No ADS, controlled	48.66 ± 0.17	65.92 ± 0.02	74.10 ± 0.05	47.07 ± 0.04	34.96 ± 0.05
No ADS, main setting	48.99 ± 0.06	65.97 ± 0.15	74.20 ± 0.11	47.52 ± 0.04	35.62 ± 0.10
TBAPR					
Controlled	50.08 ± 0.33	66.23 ± 0.23	73.26 ± 0.34	44.66 ± 0.04	26.84 ± 0.03
Main setting	49.69 ± 0.14	66.34 ± 0.19	73.33 ± 0.08	44.81 ± 0.04	27.11 ± 0.01

C.2 Component comparisons

Table A5 reports component comparisons on AERI-PEDES under the same controlled conditions. The BFC-only row uses direct scores after calibration. Candidate selection applies both learned scorers without interval ordering. Interval ordering retains the initial first candidate and uses neighborhood scores within the remaining fixed intervals. Complete RCR applies both stages. The first-scorer control removes the second scorer but retains interval ordering. The unrestricted-tail control retains the selected first candidate but sorts all remaining candidates together, allowing changes to the top-five and top-ten sets.

Table A5. Component controls over three AERI-PEDES seeds. Values are mean and sample standard deviation in percent. No BFC and unbounded BFC retain the subsequent RCR stages.

Setting	Rank-1	Rank-5	Rank-10	mAP	mINP
No BFC	48.71 ± 0.14	65.87 ± 0.11	74.30 ± 0.11	47.30 ± 0.04	35.44 ± 0.07
Unbounded BFC	48.55 ± 0.10	65.35 ± 0.07	73.70 ± 0.02	46.66 ± 0.04	34.48 ± 0.04
BFC only	48.12 ± 0.19	65.73 ± 0.10	74.17 ± 0.06	45.59 ± 0.05	32.08 ± 0.06
Candidate selection	48.82 ± 0.02	65.73 ± 0.10	74.17 ± 0.06	45.66 ± 0.03	32.08 ± 0.05
Interval ordering	48.12 ± 0.19	65.73 ± 0.10	74.17 ± 0.06	47.14 ± 0.04	35.15 ± 0.06
Complete RCR	48.82 ± 0.02	65.73 ± 0.10	74.17 ± 0.06	47.22 ± 0.01	35.15 ± 0.06
First scorer and intervals	48.96 ± 0.05	65.73 ± 0.10	74.17 ± 0.06	47.24 ± 0.02	35.15 ± 0.06
Unrestricted tail	48.82 ± 0.02	64.40 ± 0.18	73.04 ± 0.08	47.91 ± 0.01	36.42 ± 0.04

Figure A2 shows the paired changes for all three seeds. Bounded calibration improves mAP by 0.555 ± 0.032 points relative to the unbounded residual. However, it produces 0.080 ± 0.023 points lower mAP than the control without BFC. Thus, the repeated runs support the benefit of constraining a learned residual relative to the tested unbounded alternative, but do not establish that calibration improves every metric relative to omitting it.

Complete RCR improves Rank-1 by 0.700 ± 0.200 points and mAP by 1.626 ± 0.043 points over direct ranking of the calibrated features. In contrast, adding the second scorer to the first scorer and interval ordering reduces Rank-1 by 0.141 ± 0.041 points and mAP by 0.025 ± 0.006 points. Both differences are negative in all three runs. The gain observed for the second scorer for the checkpoint used in the main experiments is therefore not consistent across these seeds.

Removing the interval boundaries increases mean mAP by 0.697 points and mINP by 1.269 points, but decreases Rank-5 by 1.330 points and Rank-10 by 1.134 points. The fixed intervals preserve these candidate sets at the cost of some freedom to improve full-gallery ordering. This comparison clarifies the tradeoff imposed by the constraint and does not imply that the constrained ranking dominates an unrestricted ranking on all metrics.

D Additional ranking analyses

D.1 Distribution of changes across queries

The following analyses use the main model checkpoint, with initial scores held fixed before and after RCR. Figure A3 shows the empirical cumulative distribution of the AP difference for every

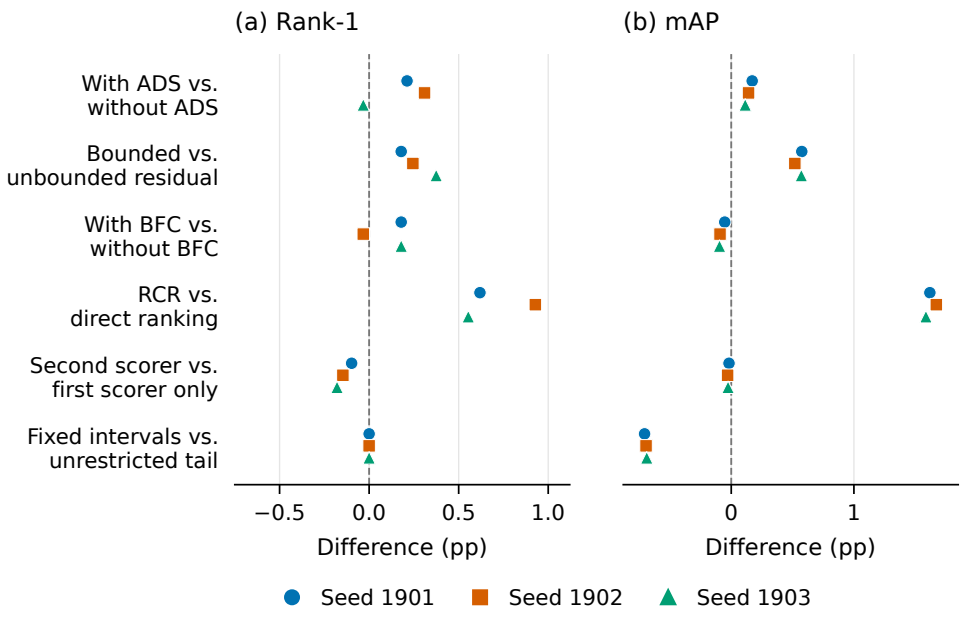


Figure A2. Paired changes across three AERI-PEDES training seeds under fixed evaluation conditions. Each point gives the metric for the first condition minus that for the second condition named on its row. The panels show Rank-1 and mAP differences in percentage points. Every seed is displayed separately.

test query. Positive values indicate improvement. The full observed range is retained, including queries with substantial degradation.

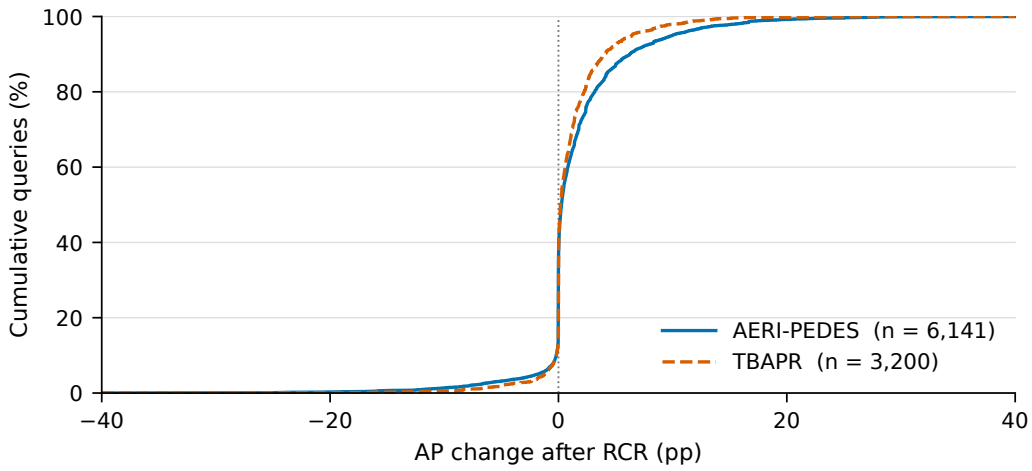


Figure A3. Empirical cumulative distributions of per-query AP changes after RCR. Both conditions use the same initial scores for all 6,141 AERI-PEDES and 3,200 TBAPR queries. The horizontal axis shows changes in percentage points, and the vertical axis shows the cumulative percentage of queries. The distribution includes improved, unchanged, and degraded results.

Using the same tolerance as the main query analysis, we treat changes of at most 10^{-4} percentage points in magnitude as unchanged. AP increases for 4,009 AERI-PEDES queries, is unchanged for 1,179, and decreases for 953. The corresponding TBAPR counts are 2,141, 477, and 582. The median changes are 0.264 and 0.175 percentage points, respectively, while the means are 1.596 and

1.078 points. The difference between each mean and median reflects the asymmetric distribution of gains. The overall improvement therefore does not imply that every query benefits from gallery context.

Table A6. Mean changes after RCR grouped by the number of matching gallery instances. Changes are in percentage points with the same initial scores. Group sizes refer to queries.

Dataset	Matching instances	Queries	ΔAP	ΔINP
AERI-PEDES	1 to 2	440	+0.44	+1.42
AERI-PEDES	3 to 4	1,592	+1.41	+2.78
AERI-PEDES	5 to 8	3,955	+1.80	+3.19
AERI-PEDES	9+	154	+1.54	+2.70
TBAPR	1 to 2	137	+0.67	+0.72
TBAPR	3 to 4	164	+0.17	+1.52
TBAPR	5 to 8	1,598	+1.11	+2.37
TBAPR	9+	1,301	+1.19	+1.42

Table A6 groups queries by the number of matching gallery instances. Mean AP and INP improve in every reported group, but the magnitude is not monotonic in the number of matches. For example, AERI-PEDES queries with five to eight matches gain 1.80 AP points on average, compared with 0.44 points for queries with one or two matches. On TBAPR, the group with three or four matches has a smaller AP gain than either adjacent group. These are descriptive comparisons of the released test sets. The number of matching gallery images per query is not independently controlled and can be associated with other differences in query difficulty.

D.2 Paired AP and INP distributions

Figure A4 compares the values before and after RCR for every test query. Each panel uses the complete range from 0 to 100 percent, and the same logarithmic count scale is used throughout. Binning reduces overlap between queries with similar values without omitting the queries at either endpoint. The diagonal provides a reference for comparing the paired values.

Mean AP increases by 1.596 points on AERI-PEDES and 1.078 points on TBAPR. Mean INP increases by 2.942 and 1.871 points, respectively. Points below the diagonal correspond to queries with decreased AP or INP, consistent with the preceding analysis. AP reflects the positions of all relevant instances, whereas INP depends on the last relevant instance. A later first relevant match can therefore reduce AP even when the last relevant match moves earlier. The qualitative examples in Section F.3 illustrate this distinction.

D.3 Changes in Rank-1 retrieval accuracy

Figure A5 shows changes in the correctness of the top-ranked result. On AERI-PEDES, RCR changes this result from incorrect to correct for 200 queries and from correct to incorrect for 162 queries, giving a net increase of 38 correct results. On TBAPR, the corresponding counts are 79 and 74, giving a net increase of five. These differences yield Rank-1 increases of 0.619 and 0.156 percentage points when the same initial scores are used.

Most queries retain the correctness of their first result, while both types of transition occur on each benchmark. Reporting only the net Rank-1 change would conceal the number of queries whose top-ranked result becomes incorrect. The interval constraint provides a separate guarantee on the top-five and top-ten candidate sets, but it does not protect the first result from an incorrect promotion. Identity labels are used only to compute these transitions after retrieval.

E Computational cost

E.1 Measurement protocol

We measure retrieval cost on an NVIDIA GeForce RTX 5090 with 32 GB of device memory and an Intel Xeon Gold 6459C CPU. The environment uses PyTorch 2.7.0, CUDA 12.8, and driver 580.76.05. The GPU is reserved for the measured process. The encoder uses BF16 autocasting, while normalization, similarities, calibration, and context calculations use FP32. TF32 is disabled.

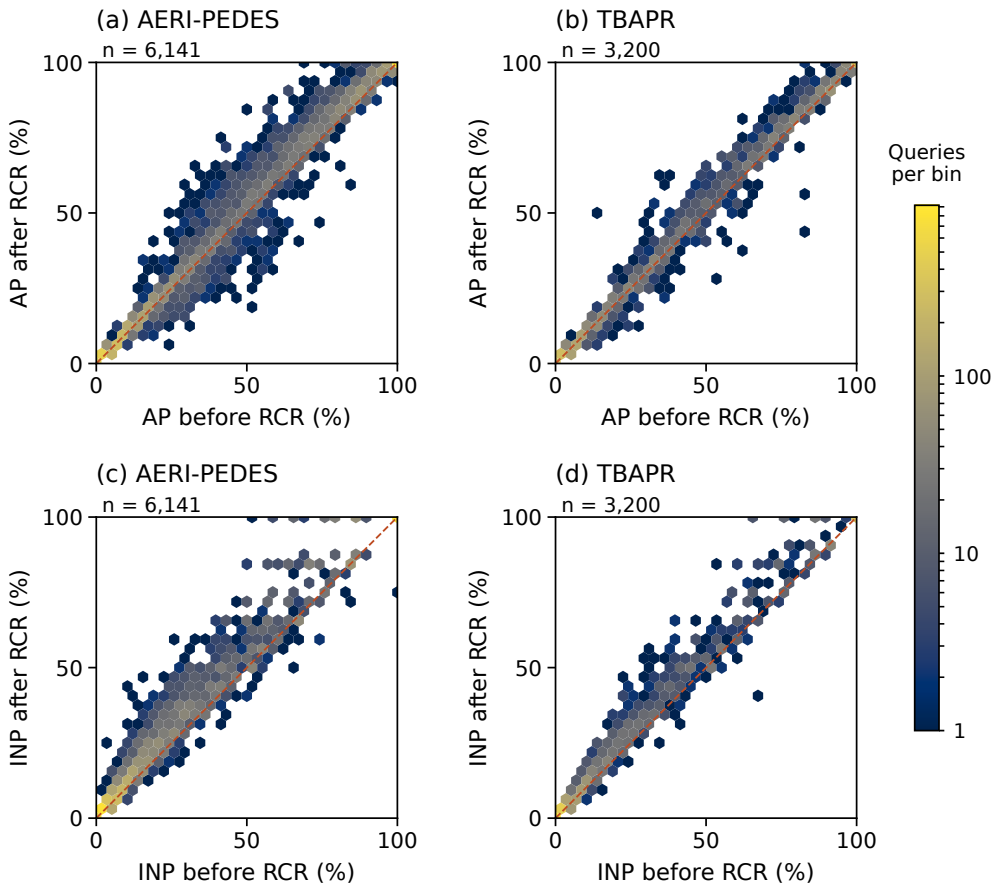


Figure A4. Paired retrieval quality before and after RCR with the same initial scores. Panels (a) and (b) show AP, and panels (c) and (d) show INP for AERI-PEDES and TBAPR, respectively. Hexagonal bins count queries on a shared logarithmic color scale. The dashed diagonal denotes equal values. All 6,141 AERI-PEDES queries and 3,200 TBAPR queries are included in each corresponding panel.

Online measurements use batch size 1 and include tokenization of the input caption, host-to-device transfer, text encoding, full-gallery ranking, and transfer of the output ranks to the CPU. Gallery representations are cached, but query embeddings are not. We perform 50 warm-up query evaluations and measure three complete passes over every test query. The reported mean and 95th percentile use all measured queries across the three passes. They describe execution latency on this system rather than uncertainty across training runs.

Offline gallery preparation uses batch size 64 and includes image decoding, resizing, encoding, calibration, and construction of the gallery graph. We measure three complete gallery passes after 50 warm-up iterations with image batches, using a warm filesystem cache. This measurement excludes encoder training and auxiliary description generation. Cache size includes the required FP32 gallery features and graph. It excludes model parameters, query embeddings, full score matrices, identity labels, and ground-view images.

E.2 Measured cost and scaling

Table A7 separates the original CFAN checkpoint, the adapted encoder before BFC and RCR, and the evaluated BCR configurations. Under the main evaluation setting, mean query latency is 8.03 ms on AERI-PEDES and 7.98 ms on TBAPR. The table reports gallery preparation, query latency, and memory separately to distinguish offline preparation from per-query computation.

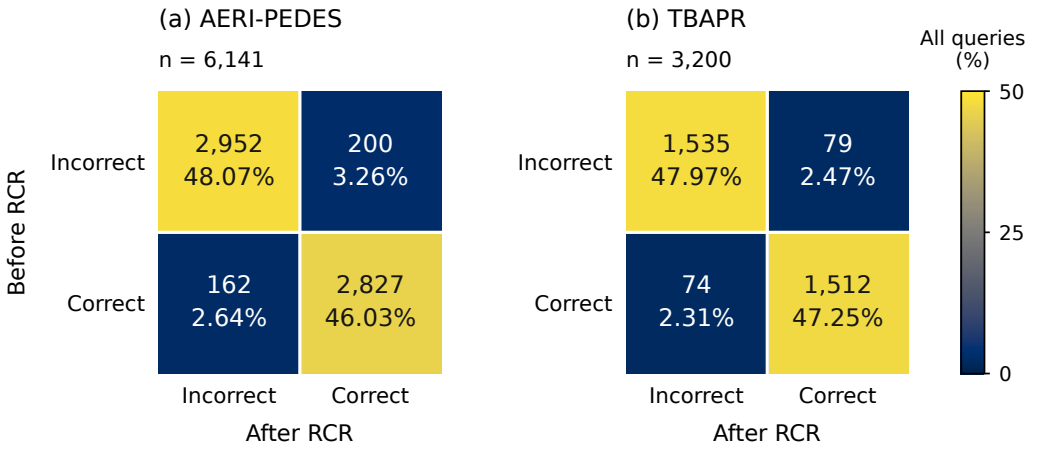


Figure A5. Transitions in the correctness of the first retrieved instance after RCR. Rows denote the result before RCR, and columns denote the result after RCR. Each cell reports a query count and its percentage of the complete test set. Percentages use all queries in the panel as the denominator, rather than the number in a row. Colors share the same scale across datasets.

Table A7. Retrieval cost under the stated measurement protocol. Gallery preparation is in seconds, query latency in milliseconds, and memory in MiB. GPU memory is the peak allocated memory during query evaluation.

System	Gallery	Mean	95th pct.	Cache	GPU
AERI-PEDES					
CFAN	2.36	6.75	8.11	10.79	676.43
Adapted encoder	2.36	6.68	6.76	10.79	678.93
BCR, controlled	2.39	7.92	8.12	11.07	680.21
BCR, main setting	4.24	8.03	9.09	21.86	691.00
TBAPR					
CFAN	1.55	6.67	6.82	6.25	673.40
Adapted encoder	1.55	6.69	7.18	6.25	673.40
BCR, controlled	1.53	7.85	8.40	6.41	674.57
BCR, main setting	2.58	7.98	8.83	12.66	680.82

BFC and the two scorers add $512^2 + 14 + 14 = 262{,}172$ learned parameters after the encoder. This count does not describe all trainable parameters in the method because encoder adaptation updates the full model. For a gallery of size N and embedding dimension D, direct construction of the gallery graph costs $O(N^2D)$. A query requires $O(ND)$ similarity computation, $O(Nk)$ neighborhood aggregation with $k \leq 5$, and $O(N \log N)$ sorting. The graph is reused across queries.

The measurements cover retrieval with the trained model checkpoints. They do not provide a complete accounting of pretraining, model selection, or description-generation cost. The accompanying result files document the measurement scope, summary statistics, and fixed configurations used for these comparisons.

F QUALITATIVE ANALYSIS

F.1 TRAINING IMAGES AND AUXILIARY DESCRIPTIONS

Figure A6 shows one training record from each of the three source datasets represented in AERI-PEDES. The released and auxiliary descriptions are reproduced verbatim from the stored training records. Ground-view images are shown to provide visual context. Only the aerial image is supplied to the generation model. These examples are the first stored record from each source and are not selected using retrieval performance.

(a) AG-ReID.v2 — Training identity 3316

Ground



Aerial



Released description

A middle-aged man wearing a blue-and-white plaid short-sleeve shirt, gray sneakers, and black shorts.

Cached auxiliary description

Person wears light short-sleeved shirt, dark shorts, and dark shoes with white soles. No carried items visible. View from above shows back and legs on paved ground.

(b) G2APS-ReID — Training identity 2532

Ground



Aerial



Released description

Red T-shirt and red skirt for a normal brunette.

Cached auxiliary description

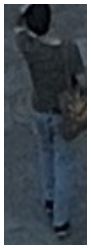
Person wears a dark patterned top and light-colored pants. Footwear is indistinct. No carried objects are visible due to low resolution and image quality.

(c) AG-VPReID — Training identity 42

Ground



Aerial



Released description

The adult male pedestrian, who was walking, had an average height, a slim body shape, and was dressed in a dark short-sleeved T-shirt paired with light blue jeans, complemented by black sneakers and a brown shoulder bag, and he sported short and dark hair.

Cached auxiliary description

Person wears dark top with light collar, light blue jeans, dark footwear. Carries a large, patterned bag over one shoulder. Image is grainy and low-resolution, limiting detail visibility.

Figure A6. Real training images and their associated descriptions from (a) AG-ReID.v2, (b) G2APS-ReID, and (c) AG-VPReID within AERI-PEDES. Each panel shows a ground-view image, an aerial image, the released annotation, and the fixed auxiliary description used by ADS. Images retain their original aspect ratios. The auxiliary text is a model output and is not treated as a verified annotation.

The descriptions differ in the attributes they specify. In panel (a), the auxiliary text uses a broad description of the shirt instead of the plaid pattern in the released annotation. Panel (b) contains a disagreement about garment type and color. Panel (c) retains a description of the carried bag while expressing uncertainty about image detail. Some cached text also includes statements about the background or image quality despite the generation instructions. These examples illustrate the noise that can remain in automatically generated supervision. They do not measure the frequency or accuracy of individual attributes across the cache.

F.2 RETRIEVAL IMPROVEMENTS

Figures A7 and A8 show cases with improved and decreased AP, respectively, using the checkpoint from the main experiments. Each comparison uses identical initial scores before and after RCR. Each row shows the five highest-ranked gallery images in retrieval order. Query text is displayed as a verbatim excerpt for readability, while retrieval uses the complete released description. AP is computed over the full gallery. Blue borders and the letter M indicate an identity match, and orange borders and N indicate a nonmatch. Identity labels are used only for this evaluation and visualization.

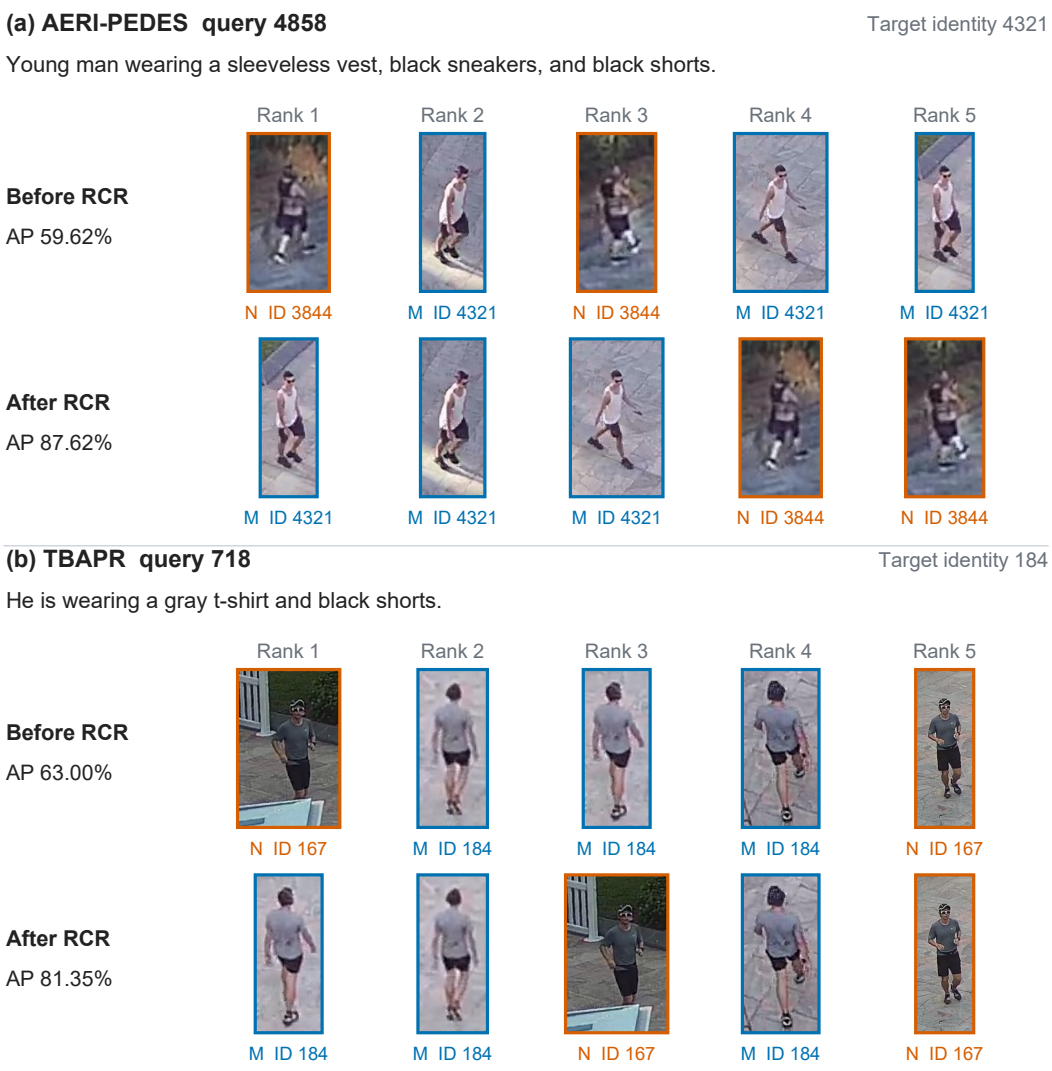


Figure A7. Selected retrieval improvements on (a) AERI-PEDES and (b) TBAPR. Each pair compares the rankings before and after RCR for the same query. AP is measured over the full gallery. The top-five candidate set is unchanged, and M and N denote matching and nonmatching identities. These examples are selected to illustrate improvements rather than sampled at random.

For AERI-PEDES query 4858, the five relevant instances move from ranks 2, 4, 5, 6, and 7 to ranks 1, 2, 3, 6, and 7. AP increases from 59.62 to 87.62 percent, while INP remains 71.43 percent because the last relevant rank is unchanged. For TBAPR query 718, a relevant instance replaces the incorrect first result, and the last relevant instance moves from rank 9 to rank 8. AP increases from 63.00 to 81.35 percent, and INP increases from 66.67 to 75.00 percent. Both examples retain the number of relevant instances within the first five ranks.

F.3 FAILURE CASES

Figure A8 shows the largest AP decrease on each benchmark among queries whose first result changes from correct to incorrect. This selection deliberately examines severe failures. It does not estimate their frequency, which is reported using all test queries in Sections D.1 and D.3.



Figure A8. Selected failures on (a) AERI-PEDES and (b) TBAPR. Each pair retains the same top-five candidates, but an incorrect candidate is promoted to rank 1. AP is measured over the full gallery. Labels M and N denote matching and nonmatching identities. Case selection is specified in the accompanying text.

For AERI-PEDES query 4519, the two relevant ranks change from 1 and 45 to 3 and 37. AP falls from 52.22 to 19.37 percent even though INP increases from 4.44 to 5.41 percent. The AP gain from moving the last relevant image earlier is outweighed by the loss from moving the first relevant image from rank 1 to rank 3. For TBAPR query 1593, the two relevant ranks change from 1 and 3 to 2 and 5. AP falls from 83.33 to 45.00 percent, and INP falls from 66.67 to 40.00 percent.

Incorrect candidates share broad clothing attributes with the query, but visual similarity does not establish identity. The examples show incorrect candidate selection but do not identify the contribution of each context coordinate. Preserving the candidate sets in Proposition B.2 therefore does not guarantee improved retrieval accuracy.