

Hybrid Token Optimization for Text-to-Image Person Re-identification

Baohua Qiang

Guilin University of Electronic Technology

Haoran Chen

Guilin University of Electronic Technology

Ruidong Chen

pgezcrb@163.com

Inner Mongolia Normal University

Shihao Zhang

Guilin University of Electronic Technology

Zhengli Xu

Guilin University of Electronic Technology

Wenhui Zhang

Guilin University of Electronic Technology

Qingfan Deng

Guilin University of Electronic Technology

Research Article

Keywords: Text-to-image person re-identification, cross-modal retrieval, representation learning, token optimization, dictionary learning, token selection

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11199036/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Hybrid Token Optimization for Text-to-Image Person Re-identification

Baohua Qiang¹, Haoran Chen¹, Ruidong Chen^{2*},
Shihao Zhang^{1*}, Zhengli Xu¹, Wenhui Zhang¹, Qingfan Deng¹

¹ Guangxi University Engineering Research Center of Cloud Network
Convergence and Data Security, Guilin University of Electronic
Technology, Guilin, 541004, China.

² College of Computer Science and Technology, Inner Mongolia Normal
University, Hohhot, 010022, China.

*Corresponding author(s). E-mail(s): pgezcrb@163.com;
zhangshihao@guet.edu.cn;

Contributing authors: qiangbh@guet.edu.cn; chr5@mails.guet.edu.cn;
zhenglixu2011@163.com; zhangwh@guet.edu.cn;
2276406237@mails.guet.edu.cn;

Abstract

Text-to-image person re-identification (TIReID) enables natural-language search over pedestrian image collections. Although dual-grained methods combine global and local matching, recent dual-grained pipelines use encoder-produced global tokens for coarse matching and attention-ranked local tokens for fine-grained matching. This interface leaves global representations without complementary priors, local evidence dependent on a single ranking cue, and image-text sequence asymmetry insufficiently addressed. To address these limitations, we propose Hybrid Token Optimization (HTO), a parameter-efficient, role-aware framework at the encoder-matcher interface. HTO comprises two coordinated components. Dictionary Prior Enhancement uses modality-specific dictionaries, cross-attention, and conservative residual blending to refine global-token representations. Spatial-Energy Heterogeneous Token Selection combines attention-based saliency, a feature-norm cue, and an adaptive minimum-coverage constraint to broaden local-token coverage in both modalities. Both components leave the established matching heads unchanged and operate during both training and inference, while the auxiliary losses are used only for training. Under a fixed-seed protocol, HTO yields Rank-1 point-estimate gains of 0.45, 0.40, and

2.25 percentage points over the reproduced baseline on CUHK-PEDES, ICFG-PEDES, and RSTPReid, respectively. It increases the parameter count by less than 0.4% and is also compatible with Test-Time Human-Centered Interaction. These results suggest that role-aware token-interface refinement can improve fine-grained cross-modal retrieval without redesigning the matcher.

Keywords: Text-to-image person re-identification, cross-modal retrieval, representation learning, token optimization, dictionary learning, token selection

1 Introduction

Cross-modal representation learning maps visual and linguistic observations into a shared space despite their different structures and distributions (Xiong et al. 2024; Li et al. 2026). Text-to-image person re-identification (TIReID) (Li et al. 2017; Jiang and Ye 2023; Qin et al. 2024; Yan et al. 2023; Zuo et al. 2024; Deng et al. 2025; Li and Xie 2024) is a fine-grained instance that requires identity-discriminative alignment between image and text. Given a natural-language description, a TIReID model ranks a pedestrian image gallery and is expected to place images of the described identity first. Unlike image-based person re-identification, TIReID remains applicable when no probe image is available.

Existing TIReID methods have progressed from global alignment to increasingly fine-grained matching. Global methods (Zhang and Lu 2018; Zheng et al. 2020) compress an image and a sentence into single vectors, while local methods (Ji et al. 2022; Shen et al. 2023; Yan et al. 2023; Jiang and Ye 2023; Deng et al. 2025; Li and Xie 2024) establish correspondences between image patches and words to distinguish subtle attributes. Related image-text matching work also learns complementary global and local similarity representations (Li et al. 2026). More recently, dual-grained TIReID frameworks (Qin et al. 2024) have combined coarse-grained basic global embeddings (BGE) with fine-grained token selection embeddings (TSE), providing a strong balance between global ranking and local alignment. Interactive TIReID further improves incomplete queries through dialogue or test-time refinement (Bai et al. 2025; Qin et al. 2025).

Despite their strong performance, RDE-style dual-grained methods rely on encoder-produced global tokens for BGE and attention-ranked local tokens for TSE. These input roles can benefit from targeted processing at the encoder-matcher interface, particularly because image and text sequences have different structures. Three linked weaknesses remain:

- **Global-role compression bottleneck.** The single global token (the visual [CLS] and the textual [EOS]) compresses information from an entire image or sentence into one encoder representation. Such compression may leave fine discriminative details weakly represented for hard samples with similar appearances.

- **Local-role sampling blind spot.** Ranking local tokens by attention mainly reflects their relevance to the global representation. This single criterion may overlook additional high-magnitude feature responses that receive moderate attention weights but can broaden the candidate evidence available for fine-grained matching.
- **Image-text sequence asymmetry at the encoder-matcher interface.** Image patches and text words differ in structure, valid sequence length, and feature distribution. A token interface for dual-grained TIReID should accommodate these differences when enhancing global representations and selecting local evidence for the existing BGE and TSE heads.

To address these limitations, we propose Hybrid Token Optimization (HTO), a lightweight, role-aware framework that refines heterogeneous encoder outputs before dual-grained matching. HTO treats global enhancement and local selection as two coordinated responses to the role mismatch at the encoder-matcher interface. Dictionary Prior Enhancement (DPE) serves the global role by using modality-specific dictionaries, cross-attention, and conservative residual blending to refine compressed global-token representations. Spatial-Energy Heterogeneous Token Selection (SHTS) serves the local role by combining attention-based saliency with a feature-norm cue and an adaptive minimum-coverage constraint to construct broader local-token evidence. Both operate at the encoder-matcher interface without changing the BGE and TSE heads. The main contributions of this paper are as follows:

- We develop DPE, whose modality-specific learnable dictionaries are updated through task-loss gradients together with a slow cross-iteration exponential moving average (EMA) correction. Conservative residual blending further limits perturbations to the pre-trained representations.
- We develop SHTS, which supplements attention-based saliency with feature-norm ranking and adaptive minimum coverage for modality-specific local-token selection. We use BCC and OCFR (Zhang et al. 2024) as training-only auxiliary regularizers rather than as part of the selector.
- We establish a role-aware encoder-matcher interface for dual-grained TIRReID. Modality-specific dictionaries, local masks, and valid-length handling accommodate image-text sequence asymmetry while preserving the BGE and TSE heads. Fixed-seed experiments evaluate the complete framework and its compatibility with THI-based query refinement.

The remainder of this paper is organized as follows. Section 2 reviews the related work on text-to-image person re-identification and token optimization. Section 3 elaborates the proposed HTO framework, including the DPE and SHTS modules and the training objectives. Section 4 presents the experimental results and analyses on three benchmarks. Section 5 concludes this paper.

2 Related Work

HTO lies at the intersection of cross-modal representation learning and token-level optimization. Related work has preserved cross-modal similarity at multiple levels (Xiong et al. 2024), modeled bidirectional relations for text-based person

retrieval (Li and Xie 2024), and learned distinct global and local similarities for image-text matching (Li et al. 2026). Against this background, we review the three technical lines most relevant to HTO: TIReID, token enhancement, and token selection.

2.1 Text-to-Image Person Re-identification

TIReID aims to retrieve images of a target identity from a natural-language description (Li et al. 2017; Jiang and Ye 2023; Qin et al. 2024; Yan et al. 2023; Zuo et al. 2024; Deng et al. 2025; Li and Xie 2024). Existing methods can be broadly divided according to their matching granularity. Early global methods (Zhang and Lu 2018; Zheng et al. 2020) map an image and a sentence into a common embedding space. Local methods (Shao et al. 2022; Wang et al. 2021; Ji et al. 2022; Shen et al. 2023; Yan et al. 2023; Jiang and Ye 2023; Li and Xie 2024) further align image regions with words or phrases to capture fine-grained attributes. Dual-grained methods combine the two representations; in particular, RDE (Qin et al. 2024) uses BGE for coarse matching and TSE for selected-token matching.

Recent studies have improved this pipeline in several ways. Domain-specific pre-training strengthens TIReID representations (Yang et al. 2023; Tan et al. 2024; Jiang et al. 2025), uncertainty-aware and robust objectives address ambiguous or noisy correspondence (Qin et al. 2024; Deng et al. 2025), and interactive approaches refine incomplete search intent at inference (Bai et al. 2025; Qin et al. 2025). Recent text-based person retrieval work also reconstructs structured semantic cues from visual background clutter and linguistic redundancy (Chen et al. 2026). These methods improve the encoder, training objective, or query, but the encoder-output tokens entering BGE and TSE are still rarely treated as a unified interface with distinct global and local roles. HTO instead targets this intermediate stage without redesigning the encoder or matching heads and remains compatible with query refinement.

2.2 Token Enhancement

The wide adoption of Transformers (Dosovitskiy et al. 2021; Liu et al. 2021; Radford et al. 2021) has motivated token-level enhancement through external priors. Dictionary learning has long provided compact priors for signal processing and computer vision (Yang et al. 2010; Mairal et al. 2010). Deep dictionary models subsequently embedded learnable atoms into neural networks (Mohammed et al. 2022; Song et al. 2019), and ATD (Zhang et al. 2024) coupled a dictionary with Transformer cross-attention to refine full patch sequences for image super-resolution.

Among these approaches, ATD (Zhang et al. 2024) uses the full image token sequence to query a dictionary for super-resolution. DPE adopts the dictionary-prior principle for a different input role: only the global visual [CLS] or textual [EOS] token queries a non-shared modality-specific dictionary, and the retrieved prior is conservatively blended with the pre-trained representation before BGE.

2.3 Token Selection

Token selection retains informative tokens while suppressing redundant content. Generic visual methods (He et al. 2022; Rao et al. 2021; Bolya et al. 2023) perform

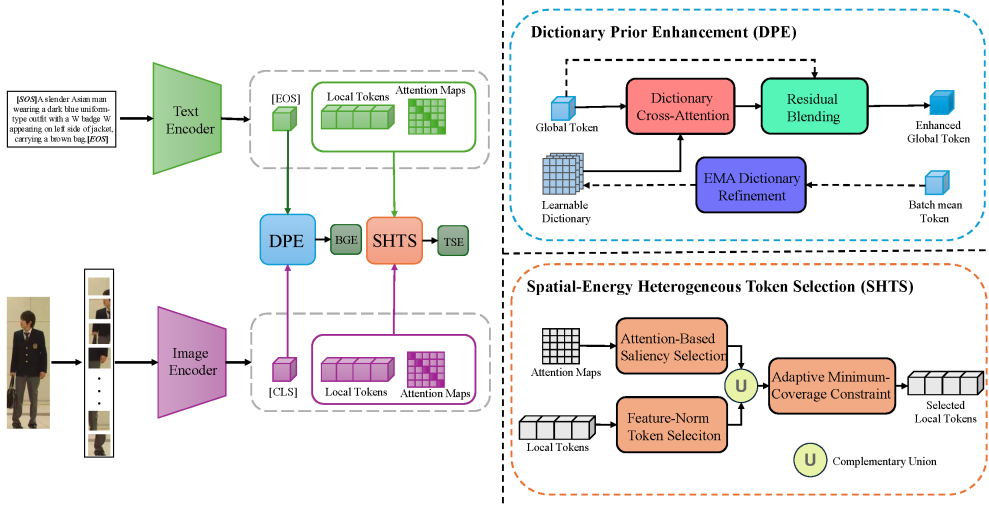


Fig. 1 Overview of HTO. DPE refines global-token representations using modality-specific learnable dictionaries that combine task-loss gradient updates with slow cross-iteration EMA correction during training and remain frozen at inference. SHTS combines attention-based saliency and feature-norm selection with an adaptive minimum-coverage constraint during both training and inference; the auxiliary losses are applied only during training. The enhanced global tokens feed BGE, whereas the selected local tokens feed TSE; their scores are combined for retrieval.

attention screening, gated sparsification, or similarity-based merging. In image-based person Re-ID, selective relation-aware representations use spatial and channel attention to emphasize salient pedestrian features (Luo et al. 2022). In image-text learning, LAPS uses linguistic supervision to identify redundant visual patches (Fu et al. 2024), while VisPruner supplements text-visual attention with visual importance and diversity (Zhang et al. 2025). TS2-Net selects spatial and temporal video tokens for text-video matching (Liu et al. 2022). Multimodal object re-identification methods further adapt selection to multiple visual streams (Yang et al. 2023; Wang et al. 2024), and EDITOR (Zhang et al. 2024) combines spatial attention with frequency analysis.

The reviewed visual-token selectors and multi-visual-stream methods motivate complementary selection criteria, but their token rules do not directly specify how to handle both image patches and text words as local inputs to a dual-grained TIREID matcher. SHTS applies attention-based saliency and a feature-norm cue to each modality independently, then uses modality-specific valid lengths and an adaptive minimum-coverage constraint before TSE. The same selection mechanism is used during training and inference.

3 Methodology

3.1 Problem Formulation

In TIReID, a natural-language description is used to retrieve the target pedestrian image from a large gallery. Let $q \in \mathcal{Q}$ denote a text query and $v \in \mathcal{V}$ a pedestrian image, where $\mathcal{Q}$ and $\mathcal{V}$ are the query set and the image gallery, and each image is associated with an identity label $y^p \in \{1, \dots, C\}$.

Following prior work (Jiang and Ye 2023; Qin et al. 2024), we adopt the CLIP visual encoder f^v and text encoder f^t to encode an image and a text into token sequences $\mathbf{V} = \{\mathbf{v}_g, \mathbf{v}_1, \dots, \mathbf{v}_{N_o}\} \in \mathbb{R}^{(N_o+1) \times d}$ and $\mathbf{T} = \{\mathbf{t}_s, \mathbf{t}_1, \dots, \mathbf{t}_{N_o}, \mathbf{t}_e\} \in \mathbb{R}^{(N_o+2) \times d}$, respectively. Here, $\mathbf{v}_g$ and $\mathbf{t}_e$ are the global tokens, $\mathbf{t}_s$ is the leading [SOS] token of the text sequence, and d is the feature dimension. The H -head self-attention maps from the last layer of the two encoders are denoted as $\mathbf{A}^v$ and $\mathbf{A}^t$. We further introduce adaptive token dictionaries $\mathbf{D}^v, \mathbf{D}^t \in \mathbb{R}^{M \times d}$, one for each modality, to enhance the global representation, where M is the dictionary size.

As illustrated in Fig. 1, HTO introduces role-aware heterogeneous token refinement at the encoder–matcher interface. It keeps the dual-grained matching heads unchanged and refines their inputs according to their distinct representational demands: DPE incorporates dictionary priors into the compressed global-token representations used by BGE, whereas SHTS broadens the local-token evidence used by TSE. The visual and textual paths follow the same interface principle but retain modality-specific parameters, masks, and valid lengths.

3.2 Dictionary Prior Enhancement

The BGE branch directly uses the cosine similarity between the global tokens (the visual [CLS] feature $\mathbf{v}_g$ and the textual [EOS] feature $\mathbf{t}_e$) as the coarse-grained matching signal. This design is efficient, but each token compresses an entire image or sentence into one encoder representation and may leave fine discriminative details weakly represented.

Inspired by token dictionary learning (Zhang et al. 2024), DPE adapts dictionary attention to the global role of dual-grained TIReID rather than refining a full dense token sequence. It uses each modality’s global token as the sole cross-attention query, combines the retrieved prior with conservative residual blending, and adapts non-shared dictionaries through task-loss gradients and slow cross-iteration EMA correction.

3.2.1 Dictionary Cross-Attention

Given a global token $\mathbf{x} \in \mathbb{R}^{1 \times d}$ ($\mathbf{v}_g$ for an image and $\mathbf{t}_e$ for a text) and a dictionary $\mathbf{D} \in \mathbb{R}^{M \times d}$, DPE generates the query, keys, and values through linear projections:

$$\mathbf{Q} = \mathbf{x}\mathbf{W}^Q, \quad \mathbf{K}_D = \mathbf{D}\mathbf{W}^K, \quad \mathbf{V}_D = \mathbf{D}\mathbf{W}^V, \quad (1)$$

where $\mathbf{W}^Q, \mathbf{W}^K \in \mathbb{R}^{d \times d_r}$ and $\mathbf{W}^V \in \mathbb{R}^{d \times d}$ are projection matrices and $d_r \ll d$. Atom-specific attention logits are then computed using normalized cosine similarity:

$$\begin{aligned} s_i &= \text{Sim}_{\cos}(\mathbf{Q}, \mathbf{K}_{D,i}), \quad \eta_i = \text{clamp}(\tilde{\eta}_i, 0, 1), \quad \tau_i = 1 + \eta_i \log M, \\ A_i &= \frac{\exp(\tau_i s_i)}{\sum_{j=1}^M \exp(\tau_j s_j)}, \quad i = 1, \dots, M, \end{aligned} \quad (2)$$

where $\tilde{\eta}_i$ is a learnable parameter associated with the i -th atom, and $\text{clamp}(\cdot)$ restricts its effective value to $[0, 1]$. Thus, each atom has an independently controlled logit scale τ_i . The attended prior is $\hat{\mathbf{x}} = \mathbf{A}\mathbf{V}_D$, which is conservatively blended with the original representation:

$$\alpha = \sigma(w_\alpha), \quad \tilde{\mathbf{x}} = (1 - \alpha) \cdot \mathbf{x} + \alpha \cdot \hat{\mathbf{x}}, \quad (3)$$

where $\sigma(\cdot)$ constrains the learnable coefficient to $(0, 1)$. We initialize $w_\alpha = -1$ ($\alpha \approx 0.269$) so that the pre-trained feature dominates early optimization, and learn independent coefficients for the two modalities.

3.2.2 Adaptive Dictionary Refinement

A fixed dictionary cannot track changes in the feature distribution encountered during training. We therefore retain each dictionary as a learnable parameter bank and supplement its task-loss gradient updates with a cross-iteration EMA correction. For a mini-batch, each atom is first refined as

$$\rho_i = \text{sigmoid}(\gamma_i), \quad \hat{\mathbf{D}}_i^{(t)} = \rho_i \mathbf{D}_i^{(t)} + (1 - \rho_i) \bar{\mathbf{x}}^{(t)}, \quad i = 1, \dots, M, \quad (4)$$

where $\bar{\mathbf{x}}^{(t)} = \frac{1}{B} \sum_{b=1}^B \mathbf{x}_b^{(t)}$ is the batch-mean global token and γ_i is a learnable retention logit for the i -th atom. We initialize $\gamma_i = 0$, giving $\rho_i = 0.5$ at the start of training. This batch-dependent correction is accumulated as

$$\mathbf{D}_i^{(t+\frac{1}{2})} = \beta \mathbf{D}_i^{(t)} + (1 - \beta) \hat{\mathbf{D}}_i^{(t)}, \quad (5)$$

where $\beta = 0.99$ makes the correction deliberately slow. The current mini-batch uses $\mathbf{D}^{(t)}$ to compute dictionary attention and the enhanced tokens. The EMA half-step is then applied without gradient tracking inside the same forward call. Adam subsequently applies task-loss gradients to $\mathbf{D}^{(t+\frac{1}{2})}$. EMA therefore provides a batch-level distributional correction, while atom-specific gradients from dictionary attention preserve distinct updates for individual atoms.

Because visual and textual features follow different distributions, we maintain separate dictionaries $\mathbf{D}^v$ and $\mathbf{D}^t$ with non-shared parameters. The enhanced features $\tilde{\mathbf{v}}_g$ and $\tilde{\mathbf{t}}_e$ then produce the BGE similarity $S^b = \tilde{\mathbf{v}}_g^\top \tilde{\mathbf{t}}_e / (\|\tilde{\mathbf{v}}_g\| \|\tilde{\mathbf{t}}_e\|)$.

3.3 Spatial-Energy Heterogeneous Token Selection

In the TSE branch, existing methods (Qin et al. 2024) rank local tokens by the attention weights of the last Transformer layer and select the top- k tokens for fine-grained

embedding. However, relying on a single attention criterion may overlook locally informative tokens with moderate attention responses.

Inspired by dual-domain token selection (Zhang et al. 2024), SHTS combines attention-based saliency with a feature-norm cue shared across modalities. Here, the spatial domain refers to attention-guided token localization, whereas the energy domain refers to token-response magnitude measured by the feature norm. Image and text tokens are selected independently because their sequence structures and lengths differ, and an adaptive minimum-coverage constraint prevents the selected subset from becoming too small. The resulting token subsets are used to construct TSE representations during both training and inference.

3.3.1 Attention-Based Saliency Selection

For clarity, we describe the visual side with N_o local tokens as an example; the text sequence with N_o tokens is processed analogously, with adjustments for its valid length. This branch follows the attention-based token-ranking paradigm to locate tokens that receive high responses from the global representation. Given the H -head attention maps from the last Transformer layer, we extract, within each head, the attention weights $\mathbf{a}^{(h)} \in \mathbb{R}^{N_o}$ ($h = 1, \dots, H$) from the global token (the visual [CLS] or the textual [EOS]) to the local tokens, select the s highest-weighted tokens per head, and take the union over heads to form the salient subset:

$$\mathcal{K}^s = \bigcup_{h=1}^H \text{Top-}s(\mathbf{a}^{(h)}), \quad \mathbf{m}^s[j] = \mathbb{I}[j \in \mathcal{K}^s], \quad (6)$$

where $\mathcal{K}^s \subseteq \{1, \dots, N_o\}$ is the attention-based saliency index set, $\mathbf{m}^s \in \{0, 1\}^{N_o}$ is the corresponding selection mask, j is the local-token index, and $\mathbb{I}[\cdot]$ is the indicator function. Per-head selection followed by a union provides broader attention coverage than a single aggregated attention map. We use the modality-neutral term *saliency selection* because textual tokens have no two-dimensional spatial layout.

3.3.2 Feature-Norm Token Selection

Attention measures the relevance of each local token to the global token but may overlook strongly activated local features. We therefore use the token ℓ_2 norm as an attention-independent score shared across modalities:

$$p_j = \|\mathbf{v}_j\|_2, \quad j \in \{1, \dots, N_o\}, \quad (7)$$

where p_j is the feature-norm score of the j -th token. To ensure that feature-norm selection complements rather than duplicates the saliency criterion, we select the φ highest-norm tokens from those not chosen by attention-based saliency:

$$\mathcal{K}^\varphi = \text{Top-}\varphi_{j \notin \mathcal{K}^s}(p_j), \quad \mathbf{m}^\varphi[j] = \mathbb{I}[j \in \mathcal{K}^\varphi], \quad (8)$$

where $\mathcal{K}^\varphi \subseteq \{1, \dots, N_o\}$ is the feature-norm index set. By explicitly excluding tokens already chosen by attention-based saliency, feature-norm selection adds strongly activated tokens that the attention criterion does not cover rather than duplicating the same subset.

3.3.3 Adaptive Minimum-Coverage Constraint and Embedding

Taking the union of the attention-based saliency and feature-norm selections yields the hybrid subset:

$$\mathcal{K}^h = \mathcal{K}^s \cup \mathcal{K}^\varphi. \quad (9)$$

Since the instance-dependent union $|\mathcal{K}^h|$ can occasionally be too small to provide sufficient token coverage, we impose an adaptive minimum-coverage constraint while retaining the priority of SHTS-selected tokens. Let $k = \lfloor \mathcal{R}N_o \rfloor$ be the target lower bound and $r_v = \max(\lfloor k/2 \rfloor, 20)$ the size of the visual attention supplement. We first form

$$\mathcal{K}^v = \mathcal{K}^h \cup \text{Top-}r_v(\bar{\mathbf{a}}), \quad (10)$$

where $\bar{\mathbf{a}} = \frac{1}{H} \sum_{h=1}^H \mathbf{a}^{(h)}$ is the head-averaged global attention. If $|\mathcal{K}^v| < k$, remaining tokens are added in descending attention order until k is reached; if $|\mathcal{K}^v| \geq k$, the union is retained without truncation. Hence, k is a minimum-coverage target rather than an exact output cardinality. The textual side follows the same rule with a valid-length cap and a minimum attention supplement of 10 tokens, yielding $\mathcal{K}^t$. Variable-length subsets are padded only for batched computation, while their valid lengths are preserved for pooling.

After ℓ_2 normalization, the selected visual tokens $\mathbf{V}_{\mathcal{K}^v} = \{\mathbf{v}_j\}_{j \in \mathcal{K}^v}$ are passed through the RDE embedding transformation to produce the fine-grained representation:

$$\mathbf{e}^v = \text{MaxPool}(\text{MLP}(\hat{\mathbf{V}}_{\mathcal{K}^v}) + \text{FC}(\hat{\mathbf{V}}_{\mathcal{K}^v})), \quad (11)$$

where $\hat{\mathbf{V}}_{\mathcal{K}^v} = \text{L2Norm}(\mathbf{V}_{\mathcal{K}^v})$, $\text{MLP}(\cdot)$ is a two-layer perceptron, $\text{FC}(\cdot)$ is a single fully connected layer, and $\text{MaxPool}(\cdot)$ denotes max-pooling along the token dimension, aggregating the variable-length token sequence into a single embedding vector $\mathbf{e}^v \in \mathbb{R}^d$. The selected text tokens indexed by $\mathcal{K}^t$ are processed analogously using a separate embedding transformation with non-shared parameters, yielding $\mathbf{e}^t \in \mathbb{R}^d$. The TSE similarity is finally defined as $S^t = \mathbf{e}^{v\top} \mathbf{e}^t / (\|\mathbf{e}^v\| \|\mathbf{e}^t\|)$.

3.4 Training Objectives

At the beginning of each epoch, the current model estimates the per-sample BGE and TSE losses over the training set. Following RDE (Qin et al. 2024), we apply Confident Consensus Division (CCD) to the two branch-wise loss estimates to obtain a binary consensus correspondence label $\hat{l}_i$ for each training pair. The consensus labels weight the BGE and TSE matching losses. BCC and OCFR remain training-only auxiliary regularizers; they supervise the resulting representations but do not define the SHTS selection rule used at inference.

3.4.1 Cross-Modal Matching Loss

To learn an accurate cross-modal similarity, we impose the Triplet Alignment Loss (TAL) (Qin et al. 2024) on the BGE and TSE branches separately. Unlike the conventional triplet ranking loss, TAL replaces a hardest-negative-only target with a log-exponential aggregation over all negatives, thereby improving training stability while retaining sensitivity to hard samples. For a mini-batch, the consensus-weighted matching losses are:

$$\mathcal{L}_{\text{TAL}}^b = \sum_{i=1}^B \hat{l}_i \ell_{\text{TAL}}(S_i^b, y_i^p), \quad \mathcal{L}_{\text{TAL}}^t = \sum_{i=1}^B \hat{l}_i \ell_{\text{TAL}}(S_i^t, y_i^p), \quad (12)$$

where S_i^b and S_i^t denote the BGE and TSE similarity rows associated with the i -th pair, respectively, and y_i^p is its identity label.

3.4.2 Background Consistency Constraint

Following EDITOR (Zhang et al. 2024), we apply a Background Consistency Constraint (BCC) to the discarded tokens. Here, “background” is a convenient name: discarded visual and textual content is heterogeneous, so BCC serves only as a weak matched-pair regularizer rather than an explicit claim of semantic correspondence. Let $\mathbf{m}_v^h$ and $\mathbf{m}_t^h$ denote the modality-specific saliency-norm union masks before minimum-coverage supplementation. Their inverse masks yield the mean discarded features:

$$\begin{aligned} \mathbf{f}_{\text{bg}}^v &= \frac{\sum_j (1 - \mathbf{m}_v^h[j]) \cdot \mathbf{v}_j}{\sum_j (1 - \mathbf{m}_v^h[j]) + \epsilon}, \\ \mathbf{f}_{\text{bg}}^t &= \frac{\sum_j (1 - \mathbf{m}_t^h[j]) \cdot \mathbf{t}_j}{\sum_j (1 - \mathbf{m}_t^h[j]) + \epsilon}, \end{aligned} \quad (13)$$

where $\mathbf{f}_{\text{bg}}^v, \mathbf{f}_{\text{bg}}^t \in \mathbb{R}^d$ are the background features of the visual and textual modalities, and ϵ is a numerical-stability term that prevents division by zero. The BCC loss is defined as the mean squared error between the background features of the two modalities:

$$\mathcal{L}_{\text{BCC}} = \frac{1}{B} \sum_{i=1}^B \|\mathbf{f}_{\text{bg},i}^v - \mathbf{f}_{\text{bg},i}^t\|_2^2, \quad (14)$$

where B is the number of samples in a mini-batch and the subscript i indexes the i -th image-text pair in the batch.

3.4.3 Object-Centric Feature Refinement

Because TSE pools dynamically selected subsets, embeddings of the same identity can retain different local evidence. Object-Centric Feature Refinement (OCFR) (Zhang et al. 2024) reduces this intra-class variation by pulling each embedding toward its mini-batch identity center. Let $\mathbf{e}_i$ denote a visual or textual embedding and y_i^p its

identity. For each identity y in the mini-batch, define $\mathcal{I}_y = \{i : y_i^p = y\}$ and the corresponding center $\mathbf{c}_y = \frac{1}{|\mathcal{I}_y|} \sum_{i \in \mathcal{I}_y} \mathbf{e}_i$. The loss is

$$\mathcal{L}_{\text{OCFR}} = \frac{1}{B} \sum_{i=1}^B \|\mathbf{e}_i - \mathbf{c}_{y_i^p}\|_2^2, \quad (15)$$

where $\mathbf{c}_{y_i^p}$ is the feature center of the identity to which sample i belongs. This loss is computed independently on the visual embedding $\mathbf{e}^v$ and the textual embedding $\mathbf{e}^t$, and the final OCFR loss is the average of the two.

3.4.4 Overall Objective

Combining the above terms, the overall training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{TAL}}^b + \mathcal{L}_{\text{TAL}}^t + \lambda_{\text{BCC}} \mathcal{L}_{\text{BCC}} + \lambda_{\text{OCFR}} \mathcal{L}_{\text{OCFR}}, \quad (16)$$

where λ_{BCC} and λ_{OCFR} are hyperparameters that balance the contributions of the auxiliary losses.

Algorithm 1 summarizes the complete training and retrieval procedure.

Algorithm 1 Training and retrieval of the HTO framework.

Require: Training set $\mathcal{D} = \{(v_i, q_i, y_i^p)\}_{i=1}^N$ and number of epochs N_e
Require: HTO model $\mathcal{M}(\Theta)$ with encoders f^v, f^t and dictionaries $\mathbf{D}^v, \mathbf{D}^t \in \mathbb{R}^{M \times d}$
Require: Hyperparameters $\beta, \mathcal{R}, s, \varphi, \lambda_{\text{BCC}}, \lambda_{\text{OCFR}}$
Ensure: Optimized parameters $\hat{\Theta}$ and query-specific gallery rankings $\hat{\mathcal{R}}(q)$

- 1: Initialize the encoders from the pre-trained weights and the remaining HTO components;
 set $\gamma_i^v = \gamma_i^t = 0$ and $w_\alpha^v = w_\alpha^t = -1$
- 2: **for** $e = 1$ to N_e **do**
- 3: Calculate the per-sample BGE and TSE losses over $\mathcal{D}$
- 4: Apply CCD to the two branch-wise loss estimates and obtain the consensus labels $\{\hat{l}_i\}_{i=1}^N$
- 5: **for** each mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
- 6: Extract the DPE-enhanced BGE representations and SHTS-based TSE representations of $\mathcal{B}$
- 7: During the DPE forward pass, apply the EMA half-step in Eqs. (4) and (5) without gradient tracking
- 8: Compute the BGE and TSE similarities S^b and S^t
- 9: Calculate $\mathcal{L}_{\text{total}}$ using Eq. (16) and the consensus labels
- 10: Backpropagate and update Θ with Adam
- 11: **end for**
- 12: **end for**
- 13: Freeze all parameters and dictionaries; disable EMA corrections and auxiliary losses
- 14: For each query, apply DPE and SHTS to the query and gallery samples
- 15: Compute S^b and S^t and obtain $\hat{\mathcal{R}}(q) = \text{argsort}_{v \in \mathcal{V}}^\downarrow S(q, v)$, where $S = (S^b + S^t)/2$
- 16: **return** $\hat{\Theta}$ and $\hat{\mathcal{R}}(q)$

Training and inference. At the beginning of each epoch, CCD produces the consensus labels used by the two TAL terms. During mini-batch optimization, DPE and SHTS use the same token-processing pipeline later applied at inference: the dictionaries receive task-loss gradients and slow EMA corrections, while BCC and OCFR supervise the SHTS-based representations. At inference, DPE and SHTS remain active with all parameters and dictionaries frozen; EMA updates and auxiliary losses are disabled. During retrieval, the final ranking score averages the two branch similarities, $S = (S^b + S^t)/2$, and the query-specific output is $\hat{\mathcal{R}}(q) = \text{argsort}_{v \in \mathcal{V}}^\downarrow S(q, v)$.

4 Experiments

4.1 Datasets and Evaluation Protocol

Datasets. We evaluate HTO on three public TIReID benchmarks: CUHK-PEDES (Li et al. 2017), ICFG-PEDES (Ding et al. 2021), and RSTPReid (Zhu et al. 2021). Among the three datasets, CUHK-PEDES contains the largest numbers of identities and descriptions, ICFG-PEDES provides one relatively detailed description per image, and RSTPReid is smaller and contains naturally occurring noisy correspondences. We follow the official splits used in RDE (Qin et al. 2024); Table 1 summarizes the official split statistics.

Table 1 Official split statistics of the three TIReID benchmarks. Each cell follows the format “identities / images / descriptions.” An em dash denotes an unavailable split.

Dataset	Train	Validation	Test
CUHK-PEDES	11,003 / 34,054 / 68,098	1,000 / 3,078 / 6,158	1,000 / 3,074 / 6,156
ICFG-PEDES	3,102 / 34,674 / 34,674	—	1,000 / 19,848 / 19,848
RSTPReid	3,701 / 18,505 / 37,010	200 / 1,000 / 2,000	200 / 1,000 / 2,000

Evaluation Protocol. Following prior work (Qin et al. 2024; Jiang and Ye 2023), we use Rank- K accuracy ($K = 1, 5, 10$), mean average precision (mAP), and mean inverse negative penalty (mINP). Rank-1 measures whether the first retrieved image has the target identity, while mAP and mINP characterize broader ranking quality.

4.2 Implementation Details

HTO is built on RDE (Qin et al. 2024) with CLIP-ViT-B/16 (Radford et al. 2021) visual and text encoders. We initialize the backbone with the domain-specific pre-trained weights for TIReID released by NAM (Tan et al. 2024), following the corresponding comparison group in ICL (Qin et al. 2025).

Following the public RDE implementation, we resize images to 384×128 with stride 16, set the maximum text length to 77, and retain RDE’s modality-specific augmentation pipeline. We train for 60 epochs with Adam, batch size 64, initial learning rate 10^{-5} , a learning-rate multiplier of 5.0 for the randomly initialized cross-modal

Table 2 Key hyperparameter configurations on different datasets. Except for the parameters listed below, all other hyperparameters are kept consistent across datasets.

Hyperparameter	CUHK-PEDES	ICFG-PEDES	RSTPReid
Dictionary size M	20	32	32
Feature-norm selection count φ	4	10	10
Selection ratio $\mathcal{R}$	0.6	0.5	0.5
Additional DPE parameters	0.58M	0.59M	0.59M

and classification parameter groups, weight decay 4×10^{-5} , cosine annealing, and five warm-up epochs. TAL uses a temperature of $\tau = 0.015$ and a margin of $m = 0.1$.

Each modality uses one DPE module with reduction dimension $d_r = 16$, a learnable residual logit initialized to $w_\alpha = -1$ ($\alpha \approx 0.269$), and EMA momentum 0.99. Each dictionary atom has a learnable retention logit γ_i , initialized to zero so that $\rho_i = \text{sigmoid}(\gamma_i) = 0.5$ at the start of training. These logits and the dictionaries are optimized jointly with the remaining trainable parameters; the $2M$ retention logits across the two modalities are included in the reported DPE parameter counts. SHTS selects $s = 2$ tokens from each of 12 attention heads before feature-norm supplementation and is retained during inference. The auxiliary-loss weights used only during training are $\lambda_{\text{BCC}} = 0.1$ and $\lambda_{\text{OCFR}} = 0.5$.

Because dataset sizes and description lengths differ, the dataset-specific values of M , φ , and $\mathcal{R}$ are listed in Table 2; all other settings are shared. Following ATD (Zhang et al. 2024) and EDITOR (Zhang et al. 2024), Section 4.5 reports a discrete-grid sensitivity analysis on RSTPReid.

For the compatibility analysis, we combine HTO with ICL’s Test-Time Human-Centered Interaction (THI) (Qin et al. 2025). We use Qwen2-VL-7B-Instruct (Wang et al. 2024), fine-tuned with Low-Rank Adaptation (LoRA) (Hu et al. 2022), for five interaction rounds without changing the HTO training procedure. Experiments use two NVIDIA GeForce RTX 3090 GPUs, each with 24 GB of memory. HTO adds 0.58M–0.59M parameters to the 152.8M-parameter baseline ($< 0.4\%$).

Reproducibility and Randomness. All ablation rows use the same codebase, NAM initialization, data pipeline, precision settings, and training schedule; only the components indicated in Table 4 differ. The table therefore uses our reproduced baseline (#1), whereas Table 3 otherwise quotes published results. All runs use a fixed random seed. Reported differences are point estimates and are not interpreted as evidence of statistical significance.

4.3 Comparison with State-of-the-Art Methods

We compare HTO with representative TIReID methods: IVT (Shu et al. 2022), LCR²S (Yan et al. 2023), CFine (Yan et al. 2023), RaSa (Bai et al. 2023), IRRa (Jiang and Ye 2023), TBPS (Cao et al. 2024), CFAM (Zuo et al. 2024), RDE (Qin et al. 2024), APTM (Yang et al. 2023), NAM (Tan et al. 2024), and ICL (Qin et al. 2025).

Results without THI. Under the fixed-seed protocol, HTO reaches Rank-1 values of 78.49%, 69.55%, and 71.45% on CUHK-PEDES, ICFG-PEDES, and RSTPReid,

Table 3 Performance comparison on the three benchmarks. Results marked with $\star$ use Test-Time Human-Centered Interaction (THI). IRRA^b uses the MALS (Yang et al. 2023) pre-trained backbone; NAM^d denotes the values reproduced and reported by ICL (Qin et al. 2025); CLIP-X refers to CLIP’s own text Transformer encoder. The best results are shown in **bold**, and the second-best results are underlined. An em dash denotes an unavailable or unreported result.

Method	Image Enc. Text Enc.		CUHK-PEDES					ICFG-PEDES					RSTPReid				
			R1	R5	R10	mAP	mINP	R1	R5	R10	mAP	mINP	R1	R5	R10	mAP	mINP
Without Test-Time Human-Centered Interaction (THI)																	
IVT (Shu et al. 2022)	ViT-B	BERT	65.69	85.93	91.15	60.66	—	56.04	73.60	80.22	—	—	46.70	70.00	78.80	—	—
LCR ² S (Yan et al. 2023)	RN50	BERT	67.36	84.19	89.62	59.20	—	57.93	76.08	82.40	38.21	—	54.95	76.65	84.70	40.92	—
CFine (Yan et al. 2023)	CLIP-ViT	BERT	69.57	85.93	91.15	—	—	60.83	76.55	82.42	—	—	50.55	72.50	81.60	—	—
RaSa (Bai et al. 2023)	Swin-B	BERT	76.51	90.29	94.25	69.38	—	65.28	80.40	85.12	41.29	—	66.90	86.50	91.35	52.31	—
IRRA (Jiang and Ye 2023)	CLIP-ViT	CLIP-X	73.38	89.93	93.71	66.13	50.24	63.46	80.25	85.82	38.06	7.93	60.20	81.30	88.20	47.17	25.28
TBPS (Cao et al. 2024)	CLIP-ViT	CLIP-X	73.54	88.19	92.35	65.38	49.25	65.05	80.34	85.47	39.83	7.87	62.10	81.90	87.75	48.00	25.86
CFAM (Zuo et al. 2024)	CLIP-ViT	CLIP-X	75.60	90.53	94.36	67.27	—	65.38	81.17	86.35	39.42	—	62.45	83.55	91.10	49.50	—
RDE (Qin et al. 2024)	CLIP-ViT	CLIP-X	75.94	90.14	94.12	67.56	51.44	67.68	82.47	87.36	40.06	7.87	65.35	83.95	89.90	50.88	28.08
IRRA ^a (Jiang and Ye 2023)	CLIP-ViT	CLIP-X	74.05	89.48	93.64	66.57	—	64.37	80.75	86.12	38.85	—	61.90	80.60	89.30	48.08	—
APTM (Yang et al. 2023)	Swin-B	BERT	76.53	90.04	94.15	66.91	—	68.51	82.99	87.56	41.22	—	67.50	85.70	91.45	52.56	—
NAM ^d (Tan et al. 2024)	CLIP-ViT	CLIP-X	77.47	90.84	94.67	69.43	54.08	66.76	82.02	87.17	41.45	9.53	67.15	86.55	91.90	52.00	28.46
ICL (Qin et al. 2025)	CLIP-ViT	CLIP-X	78.18	91.63	94.83	69.58	53.48	69.22	83.49	88.06	42.34	9.01	70.00	86.60	91.70	54.16	30.93
HTO	CLIP-ViT	CLIP-X	78.49	91.49	95.01	69.68	53.49	69.55	83.56	87.96	42.47	9.31	71.45	86.60	91.45	54.88	31.00
With Test-Time Human-Centered Interaction (THI)																	
ICL [*] (Qin et al. 2025)	CLIP-ViT	CLIP-X	79.06	91.26	94.72	70.44	54.70	70.05	83.35	87.91	42.70	9.13	72.55	86.60	91.30	55.19	31.72
HTO [*]	CLIP-ViT	CLIP-X	79.17	91.44	94.82	70.45	54.63	70.18	83.29	87.73	43.13	9.59	73.00	86.75	91.30	55.79	31.98

respectively. These values exceed the published ICL results by 0.31, 0.33, and 1.45 percentage points; Table 4 provides the controlled comparison with our reproduced baseline. The largest difference occurs on RSTPReid, which has shorter and noisier descriptions than the other two benchmarks.

Combined results with THI. With THI, HTO reaches Rank-1 values of 79.17%, 70.18%, and 73.00%, exceeding ICL^{*} by 0.11, 0.13, and 0.45 percentage points, respectively. In these runs, its mAP is higher by 0.43 points on ICFG-PEDES and 0.60 points on RSTPReid. This comparison shows that HTO can be combined with THI under the same evaluation setup: HTO refines the representations before matching, whereas THI refines the query at inference.

4.4 Ablation Study

Table 4 reports component ablations on the three datasets. Row #1 is our reproduced baseline with the same NAM initialization and training configuration. In this table, SHTS denotes the complete local-token configuration described in Sections 3.3 and 3.4: the hybrid selector is used during both training and inference, while BCC and OCFR are its associated training-only losses. The table therefore does not isolate every internal operation of that configuration.

From the results in Table 4, we make the following observations.

(1) Complementarity of DPE and SHTS. Combining DPE and SHTS (#4) improves Rank-1 over the reproduced baseline (#1) by 0.45, 0.40, and 2.25 percentage points on the three benchmarks. The largest gain occurs on RSTPReid: DPE and SHTS alone improve Rank-1 by 1.15 and 1.50 points, respectively, but slightly reduce mAP, whereas their combination also raises mAP by 0.41 points.

Table 4 Component ablation. SHTS denotes the complete hybrid local-token configuration, including its training-only BCC and OCFR losses; THI denotes Test-Time Human-Centered Interaction. The best results are shown in **bold**, and the second-best results are underlined.

No.	DPE	SHTS	THI	CUHK-PEDES		ICFG-PEDES		RSTPReid	
				R1	mAP	R1	mAP	R1	mAP
#1	$\times$	$\times$	$\times$	78.04	69.62	69.15	42.31	69.20	<u>54.47</u>
#2	$\checkmark$	$\times$	$\times$	<u>78.14</u>	<u>69.67</u>	<u>69.38</u>	42.49	70.35	54.12
#3	$\times$	$\checkmark$	$\times$	78.10	69.51	69.18	42.34	<u>70.70</u>	54.07
#4	$\checkmark$	$\checkmark$	$\times$	78.49	69.68	69.55	<u>42.47</u>	71.45	54.88
#5	$\times$	$\times$	$\checkmark$	<u>79.08</u>	70.37	<u>70.05</u>	42.70	71.30	<u>55.36</u>
#6	$\checkmark$	$\times$	$\checkmark$	79.06	<u>70.41</u>	70.02	43.16	<u>72.90</u>	55.12
#7	$\times$	$\checkmark$	$\checkmark$	78.72	70.24	69.69	43.04	72.45	54.94
#8	$\checkmark$	$\checkmark$	$\checkmark$	79.17	70.45	70.18	<u>43.13</u>	73.00	55.79

(2) Effect of DPE. DPE alone (#2) yields higher Rank-1 values on all three benchmarks, with the largest gain of 1.15 points on RSTPReid. This pattern is consistent with its role as a dictionary-derived prior for compressed [CLS]/[EOS] representations.

(3) Effect of the SHTS configuration. The SHTS configuration (#3) shows its largest gain over #1 on RSTPReid, where Rank-1 increases by 1.50 points. The corresponding gains on CUHK-PEDES and ICFG-PEDES are 0.06 and 0.03 points. Because this column includes the hybrid selector during training and inference together with its associated training-only auxiliary losses, these rows evaluate the complete local-token configuration rather than the feature-norm criterion in isolation.

(4) Compatibility with THI. Compared with THI alone (#5), full HTO plus THI (#8) achieves higher Rank-1 values on all three benchmarks. The largest gain is 1.70 points on RSTPReid, where mAP also increases by 0.43 points. Single-module combinations (#6 and #7) vary across metrics, and ICFG-PEDES mAP peaks at #6. The comparison supports compatibility at the full-framework level while showing that individual differences remain metric- and benchmark-dependent.

4.5 Fixed-Seed Sensitivity Analysis

We report discrete-grid sensitivity curves on RSTPReid for the key parameters of DPE, SHTS, and the auxiliary losses. Each point is obtained by independently training a complete model with the designated setting fixed throughout that run and then evaluating its checkpoint. The curves therefore compare separately trained configurations rather than post-training or test-time adjustments to a common checkpoint. RSTPReid is used because its moderate scale permits multiple comparison runs within the available computational budget.

4.5.1 Key Hyperparameters of DPE and SHTS

Fig. 2 summarizes the four core hyperparameters.

(a) Effect of dictionary size M . M controls both the number of dictionary atoms and, through the implemented $\log M$ factor in Eq. (2), the range of atom-specific

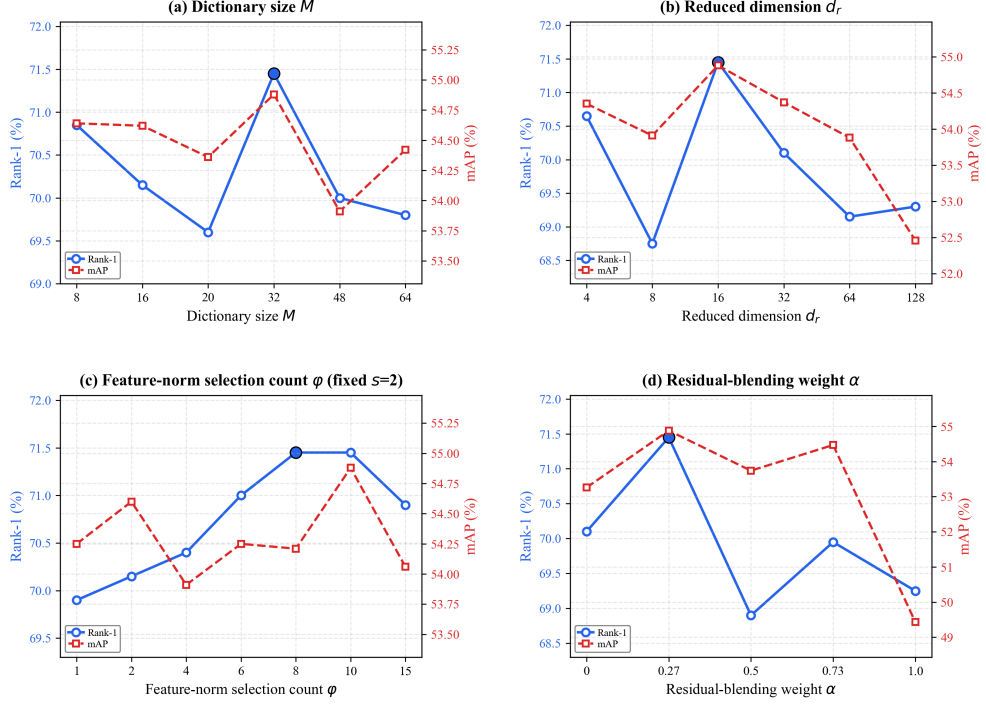


Fig. 2 Effect of the key hyperparameters of DPE and SHTS on Rank-1 and mAP (RSTPReid). (a) dictionary size M ; (b) reduction dimension d_r ; (c) feature-norm selection count φ (with $s = 2$ fixed); (d) diagnostic fixed residual-blending weight α . The left axis corresponds to Rank-1 and the right axis to mAP.

logit scaling. Fig. 2(a) therefore measures their joint implemented effect rather than dictionary capacity in isolation. The reported RSTPReid configuration uses $M = 32$, which yields the highest Rank-1 and mAP in this grid.

(b) Effect of reduction dimension d_r . d_r controls the dictionary-attention projection. Both metrics peak at $d_r = 16$ and decline as it grows.

(c) Effect of feature-norm selection count φ . φ controls feature-norm token supplementation. Rank-1 remains relatively stable for $\varphi = 8$ –10.

(d) Effect of residual-blending weight α . The final model learns $\alpha = \sigma(w_\alpha)$ from an initialization of approximately 0.269. For this diagnostic comparison, we replace the learnable coefficient with each fixed value shown in Fig. 2(d). Both metrics peak at $\alpha = 0.27$, while no injection (0) and full replacement (1) produce lower values. This pattern supports conservative initialization of the learnable blend rather than unrestricted replacement of the pre-trained representation.

4.5.2 Joint Analysis of Saliency and Feature-Norm Selection Counts

Fig. 3 jointly varies s and φ . With $s = 1$, Rank-1 varies strongly with φ and has an isolated high point at $\varphi = 6$. With $s = 4$, the curve is lower around $\varphi = 8-10$. For $s = 2$, Rank-1 remains stable at 71.45% over $\varphi = 8-10$, and mAP is highest at $\varphi = 10$; the final configuration therefore uses $s = 2$ and $\varphi = 10$.

4.5.3 Analysis of Auxiliary-Loss Weights

Fig. 4 compares the two auxiliary-loss weights. The highest point occurs at $\lambda_{\text{BCC}} = 0.1$ and $\lambda_{\text{OCFR}} = 0.5$, which are used in the final configuration.

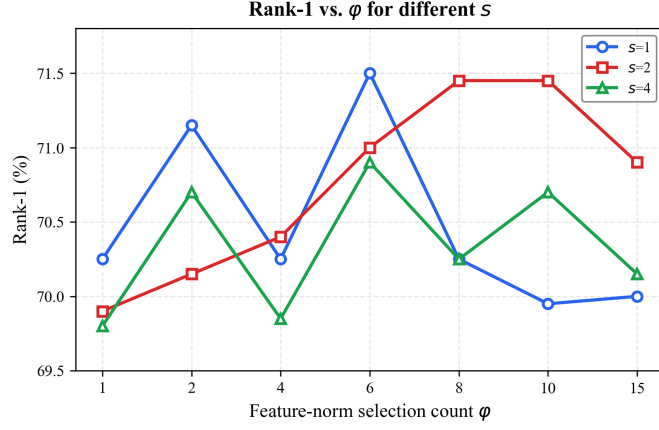


Fig. 3 Joint analysis of the per-head saliency selection count s and feature-norm selection count φ on RSTPReid.

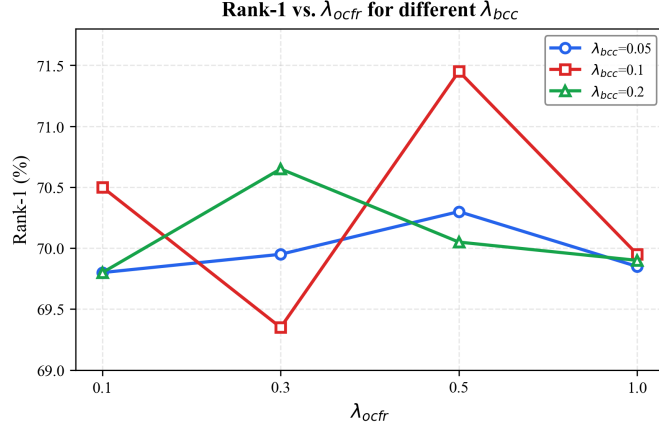


Fig. 4 Joint analysis of the BCC and OCFR loss weights on RSTPReid.

4.6 Discussion and Limitations

HTO treats global-token compression and incomplete local evidence as two manifestations of a role mismatch at the encoder–matcher interface. DPE and SHTS address the distinct evidence requirements of BGE and TSE while preserving the original matching heads and modality-specific sequence handling. Its largest observed gains occur on the smaller and noisier RSTPReid dataset, whereas the gains on the other two datasets are modest. This pattern is consistent with, but does not by itself prove, the proposed explanation that supplementary priors and broader local evidence may be more useful under the noisier benchmark conditions represented by RSTPReid. The parameter count increases by less than 0.4%, indicating parameter efficiency; however, this does not establish runtime efficiency because latency and indexing costs were not measured.

Because the evaluation uses a single fixed seed and within-dataset protocols, the results do not establish statistical significance, cross-dataset generalization, collection-scale indexing behavior, or deployed-system performance. The ablation study also does not separate the SHTS selector from its associated BCC and OCFR losses. Moreover, BCC aligns discarded visual and textual features even though these features need not be semantically equivalent. Finally, the qualitative visualization presents selected cases and does not quantify how often HTO improves retrieval. Repeated runs, finer-grained ablations, cross-domain evaluation, and system-level measurements are therefore needed.

The three benchmarks contain pedestrian images and natural-language descriptions released for research. Although HTO is evaluated only on these public benchmarks, language-guided person retrieval can be misused for intrusive surveillance or identity tracking. Any practical deployment should therefore comply with applicable privacy requirements and incorporate explicit access control, purpose limitation, and human oversight.

4.7 Retrieval Visualization

Fig. 5 compares HTO with ICL (Qin et al. 2025) on selected hard queries. HTO ranks the correct identity first in each example, illustrating that HTO can alter the ordering of visually similar candidates.

[CUHK]

She has long black hair
that goes past her shoulders
. She is wearing a white top
and is carrying a large
black and white bag .

ICL

Ours

[ICFG]

The man with short black hair
is wearing a dijon yellow
jacket and Prussian blue
bootcut pants. He is also
wearing black shoes.

ICL

Ours

[RSTP]

A man with short hair is
wearing black coat with white
writing on left arm, fitted
black pants. He is looking left
side.

ICL

Ours

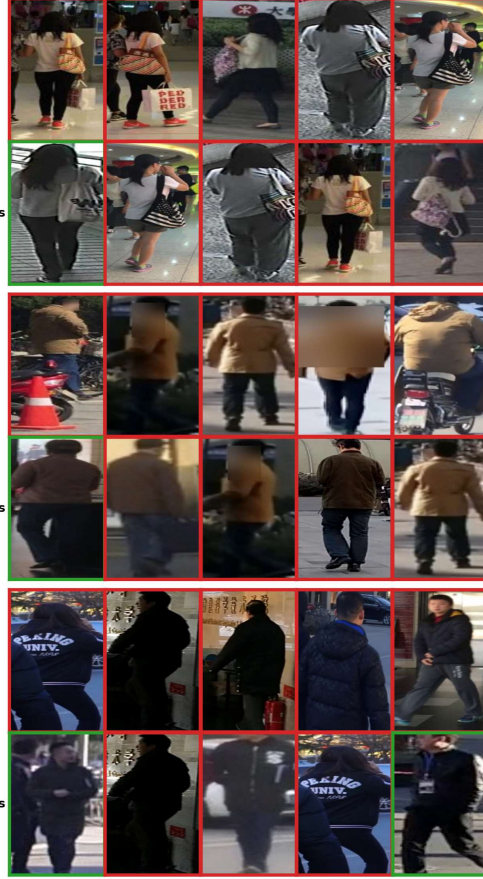


Fig. 5 Selected hard-query results for ICL (Qin et al. 2025) (top row of each group) and HTO (bottom). Green and red boxes denote correct and incorrect identities. The groups are from CUHK-PEDES, ICFG-PEDES, and RSTPReid; HTO ranks the correct identity first in each example.

5 Conclusion

This paper presents HTO, a parameter-efficient framework that addresses the role mismatch between generic encoder tokens and dual-grained matching heads in TIReID. DPE serves the global role by incorporating modality-specific dictionary priors through conservative residual blending, whereas SHTS serves the local role by combining attention-based saliency with a feature-norm cue and an adaptive minimum-coverage constraint. The resulting heterogeneous token interface preserves the original matching heads. In fixed-seed experiments on three benchmarks, HTO yields higher Rank-1 point estimates than the reproduced dual-grained baseline while increasing the parameter count by less than 0.4%. HTO is also compatible with test-time query interaction. Future work will evaluate statistical stability, conduct finer-grained ablations, investigate cross-domain transfer, and measure system-level efficiency and scalability.

Statements and Declarations

Funding. This work was supported in part by the National Natural Science Foundation of China under Grant 62262006; in part by the Guangxi Key Research and Development Program under Grants AB25069257 and AB24010112; in part by the National Natural Science Foundation of Guangxi under Grant 2025GXNSFAA069553; in part by the Guilin Science and Technology Development Program under Grant 20230110-1; in part by the Innovation Project of Guangxi Graduate Education under Grants YCBZ2025161 and YCBZ2026173; and in part by the Innovation Project of Guilin University of Electronic Technology (GUET) Graduate Education under Grants 2025YCXS050 and 2026YCXS056.

Author contributions. Conceptualization: Baohua Qiang and Haoran Chen; methodology: Baohua Qiang and Haoran Chen; software, validation, formal analysis, investigation, and visualization: Haoran Chen; data curation: Haoran Chen, Zhengli Xu, Wenhui Zhang, and Qingfan Deng; writing—original draft: Haoran Chen; writing—review and editing: Baohua Qiang, Ruidong Chen, and Shihao Zhang; supervision: Ruidong Chen and Shihao Zhang; project administration: Baohua Qiang; resources: Baohua Qiang, Ruidong Chen, and Shihao Zhang; funding acquisition: Baohua Qiang. All authors reviewed and approved the final manuscript.

Competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability. The CUHK-PEDES (Li et al. 2017), ICFG-PEDES (Ding et al. 2021), and RSTPReid (Zhu et al. 2021) datasets analyzed in this study are publicly available from their cited original sources. No new datasets were generated for this work.

References

- Bai, Y., Cao, M., Gao, D., Cao, Z., Chen, C., Fan, Z., Nie, L., Zhang, M.: RaSa: Relation and sensitivity aware representation learning for text-based person search. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 555–563 (2023). <https://doi.org/10.24963/ijcai.2023/62>
- Bolya, D., Fu, C., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your ViT but faster. In: Proceedings of the International Conference on Learning Representations (2023). <https://openreview.net/forum?id=JroZRw7Eu>
- Bai, Y., Ji, Y., Cao, M., Wang, J., Ye, M.: Chat-based person retrieval via dialogue-refined cross-modal alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3952–3962 (2025). <https://doi.org/10.1109/CVPR52734.2025.00374>

- Cao, M., Bai, Y., Zeng, Z., Ye, M., Zhang, M.: An empirical study of CLIP for text-based person search. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 465–473 (2024). <https://doi.org/10.1609/aaai.v38i1.27801>
- Chen, J., Luo, X., Kong, J., Jiang, M.: From noise to structure: structured semantic reconstruction for text-based person retrieval. *International Journal of Machine Learning and Cybernetics* **17**, 472 (2026) <https://doi.org/10.1007/s13042-026-03299-8>
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (2021). <https://openreview.net/forum?id=YicbFdNTTy>
- Deng, Y., Chen, Z., Li, C., Tang, J.: Uncertainty-aware coarse-to-fine alignment for text-image person retrieval. *Visual Intelligence* **3**, 6 (2025) <https://doi.org/10.1007/s44267-025-00078-x>
- Ding, Z., Ding, C., Shao, Z., Tao, D.: Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint* (2021) [arXiv:2107.12666](https://arxiv.org/abs/2107.12666)
- Fu, Z., Zhang, L., Xia, H., Mao, Z.: Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26307–26316 (2024). <https://doi.org/10.1109/CVPR52733.2024.02485>
- He, J., Chen, J.-N., Liu, S., Kortylewski, A., Yang, C., Bai, Y., Wang, C.: TransFG: A transformer architecture for fine-grained recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 852–860 (2022). <https://doi.org/10.1609/aaai.v36i1.19967>
- Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: Proceedings of the International Conference on Learning Representations (2022). <https://openreview.net/forum?id=nZeVKeeFYf9>
- Jiang, J., Ding, C., Tan, W., Wang, J., Tao, J., Xu, X.: Modeling thousands of human annotators for generalizable text-to-image person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9220–9230 (2025). <https://doi.org/10.1109/CVPR52734.2025.00861>
- Ji, Z., Hu, J., Liu, D., Wu, L.Y., Zhao, Y.: Asymmetric cross-scale alignment for text-based person search. *IEEE Transactions on Multimedia* **25**, 7699–7709 (2022) <https://doi.org/10.1109/TMM.2022.3225754>

- Jiang, D., Ye, M.: 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, pp. 2787–2797 (2023). <https://doi.org/10.1109/CVPR52729.2023.00273>
- Luo, X., Jiang, M., Kong, J.: Selective relation-aware representations for person re-identification. *International Journal of Machine Learning and Cybernetics* **13**, 3523–3541 (2022) <https://doi.org/10.1007/s13042-022-01610-x>
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022 (2021). <https://doi.org/10.1109/ICCV48922.2021.00986>
- Li, B., Li, Z., Wu, Y., Wei, X.: Neighborhood-interference independent graph mechanisms for image-text matching. *International Journal of Machine Learning and Cybernetics* **17**, 34 (2026) <https://doi.org/10.1007/s13042-025-02818-3>
- Li, Z., Xie, Y.: BCRA: Bidirectional cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. *Multimedia Systems* **30**(4), 177 (2024) <https://doi.org/10.1007/s00530-024-01372-2>
- Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1970–1979 (2017). <https://doi.org/10.1109/CVPR.2017.551>
- Liu, Y., Xiong, P., Xu, L., Cao, S., Jin, Q.: TS2-Net: Token shift and selection transformer for text-video retrieval. In: Proceedings of the European Conference on Computer Vision, pp. 319–335 (2022). https://doi.org/10.1007/978-3-031-19781-9_19
- Mohammed, H.J., Al-Fahdawi, S., Al-Waisy, A.S., Zebari, D.A., Ibrahim, D.A., Mohammed, M.A., Kadry, S., Kim, J.: ReID-DeePNet: A hybrid deep learning system for person re-identification. *Mathematics* **10**(19), 3530 (2022) <https://doi.org/10.3390/math10193530>
- Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research* **11**(1), 19–60 (2010)
- Qin, Y., Chen, C., Fu, Z., Peng, D., Peng, X., Hu, P.: Human-centered interactive learning via MLLMs for text-to-image person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14390–14399 (2025). <https://doi.org/10.1109/CVPR52734.2025.01342>
- Qin, Y., Chen, Y., Peng, D., Peng, X., Zhou, J.T., Hu, P.: Noisy-correspondence learning for text-to-image person re-identification. In: Proceedings of the IEEE/CVF

- Conference on Computer Vision and Pattern Recognition, pp. 27197–27206 (2024). <https://doi.org/10.1109/CVPR52733.2024.02568>
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR, Virtual (2021)
- Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.-J.: DynamicViT: Efficient vision transformers with dynamic token sparsification. In: Advances in Neural Information Processing Systems, vol. 34, pp. 13937–13949 (2021)
- Shen, F., Shu, X., Du, X., Tang, J.: Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In: Proceedings of the ACM International Conference on Multimedia, pp. 8922–8931 (2023). <https://doi.org/10.1145/3581783.3612009>
- Shu, X., Wen, W., Wu, H., Chen, K., Song, Y., Qiao, R., Ren, B., Wang, X.: See finer, see more: Implicit modality alignment for text-based person retrieval. In: Proceedings of the European Conference on Computer Vision, pp. 624–641 (2022). https://doi.org/10.1007/978-3-031-25072-9_42
- Song, J., Yang, Y., Song, Y.-Z., Xiang, T., Hospedales, T.M.: Generalizable person re-identification by domain-invariant mapping network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 719–728 (2019). <https://doi.org/10.1109/CVPR.2019.00081>
- Shao, Z., Zhang, X., Fang, M., Lin, Z., Wang, J., Ding, C.: Learning granularity-unified representations for text-to-image person re-identification. In: Proceedings of the ACM International Conference on Multimedia, pp. 5566–5574 (2022). <https://doi.org/10.1145/3503161.3548028>
- Tan, W., Ding, C., Jiang, J., Wang, F., Zhan, Y., Tao, D.: 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, pp. 17127–17137 (2024). <https://doi.org/10.1109/CVPR52733.2024.01621>
- Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. arXiv preprint (2024) [arXiv:2409.12191](https://arxiv.org/abs/2409.12191)
- Wang, C., Luo, Z., Lin, Y., Li, S.: Text-based person search via multi-granularity embedding learning. In: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 1068–1074 (2021). <https://doi.org/10.24963/ijcai.2021/148>

- Wang, Y., Liu, X., Zhang, P., Lu, H., Tu, Z., Lu, H.: TOP-ReID: Multi-spectral object re-identification with token permutation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 5758–5766 (2024). <https://doi.org/10.1609/aaai.v38i6.28388>
- Xiong, S., Pan, L., Ma, X., Hu, Q., Beckman, E.: Unsupervised deep hashing with multiple similarity preservation for cross-modal image-text retrieval. *International Journal of Machine Learning and Cybernetics* **15**, 4423–4434 (2024) <https://doi.org/10.1007/s13042-024-02154-y>
- Yang, B., Chen, J., Ye, M.: Top-k visual tokens transformer: Selecting tokens for visible-infrared person re-identification. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1–5 (2023). <https://doi.org/10.1109/ICASSP49357.2023.10097170>
- Yan, S., Dong, N., Liu, J., Zhang, L., Tang, J.: Learning comprehensive representations with richer self for text-to-image person re-identification. In: Proceedings of the ACM International Conference on Multimedia, pp. 6202–6211 (2023). <https://doi.org/10.1145/3581783.3611832>
- Yan, S., Dong, N., Zhang, L., Tang, J.: CLIP-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing* **32**, 6032–6046 (2023) <https://doi.org/10.1109/TIP.2023.3327924>
- Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. *IEEE Transactions on Image Processing* **19**(11), 2861–2873 (2010) <https://doi.org/10.1109/TIP.2010.2050625>
- Yang, S., Zhou, Y., Zheng, Z., Wang, Y., Zhu, L., Wu, Y.: Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In: Proceedings of the ACM International Conference on Multimedia, pp. 4492–4501 (2023). <https://doi.org/10.1145/3581783.3611709>
- Zhang, Q., Cheng, A., Lu, M., Zhang, R., Zhuo, Z., Cao, J., Guo, S., She, Q., Zhang, S.: Beyond text-visual attention: Exploiting visual cues for effective token pruning in VLMs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20857–20867 (2025). <https://doi.org/10.1109/ICCV51701.2025.01939>
- Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: Proceedings of the European Conference on Computer Vision, pp. 686–701 (2018). https://doi.org/10.1007/978-3-030-01246-5_42
- Zhang, L., Li, Y., Zhou, X., Zhao, X., Gu, S.: Transcending the limit of local window: Advanced super-resolution transformer with adaptive token dictionary. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2856–2865 (2024). <https://doi.org/10.1109/CVPR52733.2024.00276>

- Zhu, A., Wang, Z., Li, Y., Wan, X., Jin, J., Wang, T., Hu, F., Hua, G.: DSSL: Deep surroundings-person separation learning for text-based person retrieval. In: Proceedings of the ACM International Conference on Multimedia, pp. 209–217 (2021). <https://doi.org/10.1145/3474085.3475369>
- Zhang, P., Wang, Y., Liu, Y., Tu, Z., Lu, H.: Magic tokens: Select diverse tokens for multi-modal object re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17117–17126 (2024). <https://doi.org/10.1109/CVPR52733.2024.01620>
- Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., Shen, Y.-D.: Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications* **16**(2), 1–23 (2020) <https://doi.org/10.1145/3383184>
- Zuo, J., Zhou, H., Nie, Y., Zhang, F., Guo, T., Sang, N., Wang, Y., Gao, C.: UFineBench: Towards text-based person retrieval with ultra-fine granularity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22010–22019 (2024). <https://doi.org/10.1109/CVPR52733.2024.02078>