Reranking with Intra-modal Visual Association for Text-to-Image Person Re-Identification

Junhong Wang[1] **Changxing Ding**[1]* **Yusha Peng**[1] **Wentao Tan**[2] **Dongliang Liao**[1]
[1]South China University of Technology [2]JD Future Academy, Beijing
{eewjh,202521063538}@mail.scut.edu.cn
{chxding,liaodl}@scut.edu.cn, tanwentao1@jd.com

Abstract

Text-to-image person re-identification (TIReID) aims to retrieve images of one pedestrian using natural-language descriptions. It remains fundamentally challenging due to significant inter-modal semantic gap between sparse textual cues and fine-grained visual appearance. Existing methods mainly perform instance-wise text-image matching during inference, overlooking intra-modal visual associations among top-ranked gallery candidates, where multiple images may depict the same identity. To exploit such associations, we propose Reranking with Intra-modal Visual Association (RIVA), a plug-and-play reranking framework that leverages intra-modal visual association among retrieved candidates to guide multimodal large language model (MLLM)-based reranking. RIVA consists of two complementary components. First, we propose an MLLM-based Short-lIst Visual Association (SIVA) method, which is trained with a three-stage, dual-task curriculum to learn identity-level image-image association and then use it as visual context for text-image matching. Second, we develop a long-list visual association approach based on clustering, which clusters the top-ranked candidates in the feature space of pedestrian images, uses SIVA to examine the most likely matched cluster, and calibrates the image-text similarity scores accordingly. Extensive experiments on three popular TIReID benchmarks demonstrate that RIVA consistently improves retrieval accuracy in multiple evaluation settings. The code is available at https://github.com/Epiphany1536/RIVA.

1 Introduction

Text-to-image person re-identification (TIReID) [1–7] aims to retrieve images of a specific identity based on a textual description. It shows significant potential in a broad set of applications, such as automated video surveillance [8], social media analysis [9], and public security [10]. However, TIReID remains challenging due to significant inter-modal semantic gaps [3, 6], stemming from the inherent mismatch between the high-level semantics expressed in text and the low-level visual appearances captured in images.

To bridge this modality gap, recent studies have explored various strategies that can be categorized into two groups. The first group improves inter-modal semantic alignment by modeling cross-modal relations [6, 11, 12] or mitigating noisy correspondences [13, 14] between textual descriptions and pedestrian images. The second group mitigates the modality gap in TIReID by leveraging large-scale training datasets [7, 15–18]. However, these methods generally follow a strict instance-wise matching paradigm. They treat each gallery image as an independent retrieval target, overlooking possible intra-modal associations between gallery images [5]. In realistic TIReID scenarios, as illustrated in Figure 1, the top-K retrieved results often contain multiple images of the same pedestrian, e.g.,

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

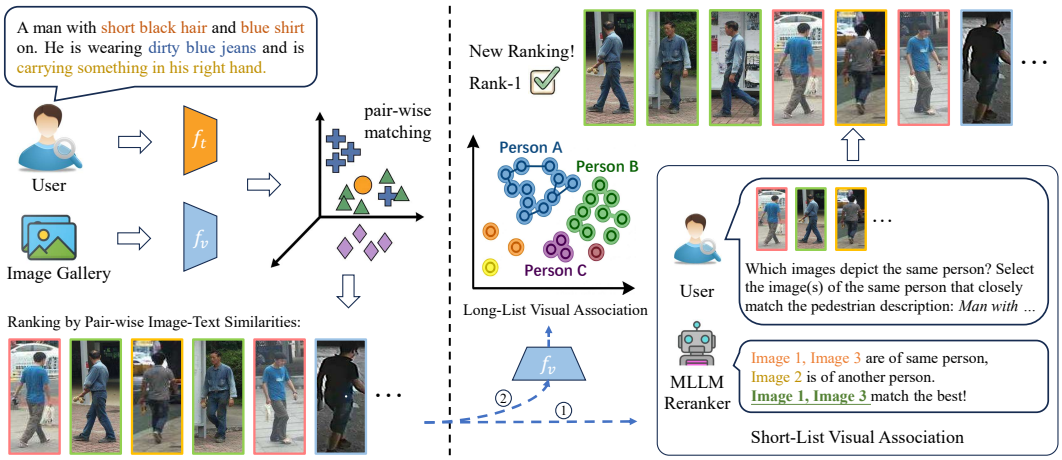

Figure 1: (Left) Existing TIReID methods, including those adopt reranking [18], generally follow a strict instance-wise matching paradigm. They compute pairwise image-text similarities and rank each gallery image independently, ignoring potential visual associations among top-ranked image candidates. (Right) Top-ranked candidates often contain images of the same identity, indicated by circles of the same color. Our RIVA sequentially employs an MLLM-based method to explore short-list visual association, and a clustering-based method to explore long-list visual association. In this way, it achieves efficient reranking over a relative large top-K image set.

consecutive video frames captured by the same camera or intermittent ones captured from different nearby cameras [18].

Consequently, a natural question arises: *can we leverage intra-modal visual associations to mitigate the inter-modal semantic gap?* Addressing this question requires joint reasoning over multiple candidate images and the text query, rather than independent image-text matching. Recent Multi-modal Large Language Models (MLLMs) make this possible by supporting multi-image reasoning, and have shown promising potential in TIReID [18]. However, their performance tends to degrade as the number of input images increases [19, 20], limiting their direct application to the fine-grained TIReID problem.

Inspired by this, we propose **RIVA** (**R**eranking with **I**ntra-modal **V**isual **A**ssociation), a reranking framework that leverages visual associations to improve cross-modal retrieval. RIVA consists of two complementary modules: an MLLM-based **S**hort-l**I**st **V**isual **A**ssociation (**SIVA**) module and a clustering-based long-list visual association module. As illustrated in Figure 1, SIVA departs from the conventional point-wise evaluation [9, 6, 13] by taking multiple candidate images as input. To endow SIVA with these capabilities, first, we design two training tasks for SIVA: the "image-image" task asks SIVA to identify which candidate images are of the same identity, enabling it to learn identity-level visual associations between visually similar images; the "text-image" task further provides a pedestrian description and asks SIVA to select the image(s) of the same pedestrian that best match the query, using the learned "image-image" relations as visual context. Second, we adopt a three-stage training pipeline. We first warm up SIVA with large-scale synthetic data [17], then adapt it to human-annotated data [9, 2, 21], and finally focus on hard samples. By incorporating such visual context into reranking, SIVA effectively reduces the image-text inter-modal gaps.

Based on SIVA, we further adopt a clustering-based approach to rerank a wider range of top-K candidates by exploiting long-list visual associations, as illustrated in Figure 1. Specifically, we first identify an anchor image using SIVA, which is regarded the most likely match to the text query within a short-list of L candidate images. We also perform clustering according to the visual features of all top-K candidates, where images in one cluster are regarded to probably have the same identity. Then, we employ SIVA to carefully examine the cluster containing the anchor image and determine which images in this cluster indeed match the text query. Finally, we promote the image-text similarity scores for the images selected by SIVA. In this way, RIVA efficiently combines both short- and long-list visual associations for wide-range TIReID reranking.

To our knowledge, this is the first study that uses visual associations among image candidates for the reranking problem in TIReID. Through a three-stage dual-task training, we enable MLLM to discriminate different identities and perform more reliable cross-modal matching within a short image-list. Through clustering-guided score calibration, we achieve reranking across a broader set of top-ranked candidates. Experimental results on three popular benchmarks demonstrate that our method significantly improves TIReID performance in multiple evaluation settings [16, 18].

2 Related Work

Text-to-Image ReID. Existing TIReID studies mainly address two key problems: how to design effective feature alignment strategies to reduce inter-modal gap, and how to construct large-scale training datasets to improve model generalization ability.

For feature alignment, early methods usually project holistic image and text representations into a shared embedding space, optimized by identity-level supervision [22, 23], adversarial learning [24], or semantic consistency constraints [9, 25, 26, 23, 27]. However, holistic matching is often insufficient for fine-grained TIReID, since different pedestrians may share similar clothing, colors, poses, or accessories, while textual descriptions are usually sparse and incomplete [1, 6]. To address this issue, later studies introduce local or part-level alignment strategies. These strategies can be roughly divided into explicit and implicit alignment methods. Explicit alignment methods rely on external or manually defined cues, such as attributes [1, 28], poses [29], body parts [29, 30], phrases [1, 29], or color-aware regions [31]. In contrast, implicit alignment methods learn fine-grained correspondences without external part annotations, using attention mechanisms [4, 32], part-aware representation learning [2, 33, 34], multi-granularity interaction [21, 3], or token-level relation reasoning [35, 6]. With the development of vision-language pre-training, recent methods further adapt pretrained VLMs, e.g., CLIP [36] and ALBEF [37], to TIReID [38–40] and model relation-aware [11, 12, 41], uncertainty-aware [42–45], prototype-based [46] or noisy-correspondence structures [13, 14]. Some works also exploit intra-modal [47, 46, 48], multi-view [5] or one-to-many relations [49] during training. Nevertheless, these structures are mainly used to regularize representation learning, while gallery candidates are still treated as isolated instances during inference.

Another line of work improves TIReID by constructing larger and more diverse training data. Since manually annotating pedestrian descriptions is expensive, several studies generate pseudo descriptions according to pedestrian attributes [7, 50, 51] and predefined templates [7, 51]. They then use the resulting large-scale image-text pairs for model training. More recent studies leverage MLLMs or dialogue-style supervision to enrich training descriptions [16], increase annotation diversity [17], reorganize low-quality texts [18], or build dialogue- and interaction-oriented ReID-domain training sets [52–54]. In addition, data filtering and curation strategies are introduced to reduce noisy or weakly matched synthetic pairs [55]. These methods improve the scale, quality, and diversity of the training data, while our work keeps the standard single-description retrieval setting and instead exploits test-time intra-modal visual associations among top-ranked gallery candidates. Therefore, our work is complementary to both existing feature alignment strategies and large-scale training data construction efforts.

MLLM-based Reranking. With the rapid development of MLLMs [56–62], recent studies have begun to adapt MLLMs to reranking, typically following a retrieve-then-rerank pipeline. One direction performs pointwise relevance estimation, where each query-candidate pair is independently evaluated [63–66, 60]; another direction adopts listwise reranking, where multiple candidates are compared jointly to capture relative preference [66, 67]. Existing multimodal rerankers have been applied to multimodal Retrieval-Augmented Generation (RAG) [68, 63], visual document reranking [64], universal multimodal retrieval [65, 66], reasoning-enhanced document retrieval [67], and unified multimodal embedding-reranking frameworks [60]. These methods demonstrate that MLLMs can serve as flexible relevance estimators for heterogeneous retrieval tasks.

Nevertheless, the above methods are mostly designed for generic image-text or document-level retrieval tasks, where relevance can often be determined from high-level semantic consistency. In comparison, TIReID is more fine-grained: visually similar pedestrians may differ only in subtle attributes, local clothing details, carried objects, or viewpoint-dependent cues [69, 13, 14]. Therefore, directly applying a generic MLLM reranker may not provide stable identity-level discrimination. To address this issue, we equip our MLLM-based SIVA model with visual association between gallery

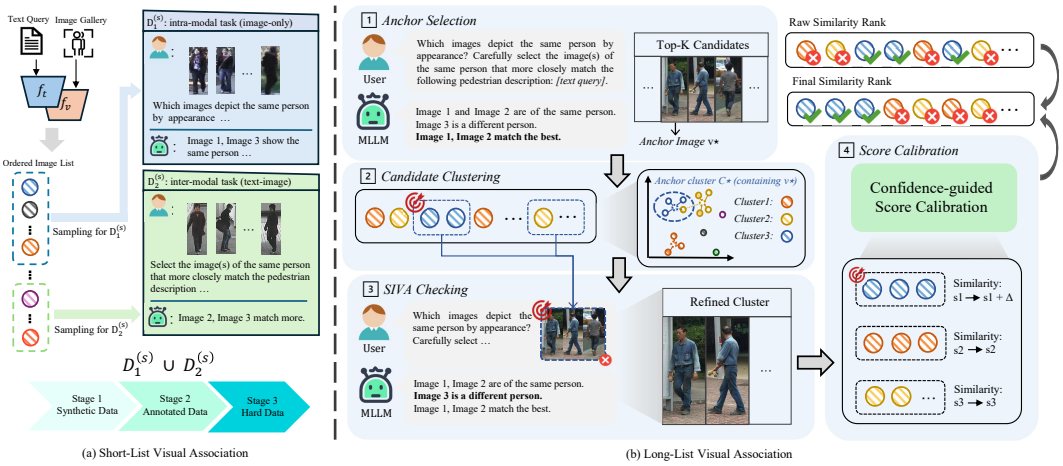

Figure 2: Overview of our RIVA framework. We explore both short- and long-list visual associations among top-K image candidates to facilitate efficient reranking. (a) We finetune an MLLM with two complementary instructions and a three-stage curriculum to learn short-list visual associations before text-image matching. (b) We adopt a clustering-based scheme to efficiently explore long-list visual associations among all top-K candidates, calibrate similarity scores for images in the anchor cluster, and finally refines the entire top-K ranking $\mathbf{V}^K(t)$.

candidates, enabling the reranking to leverage both the cross-modal reasoning ability of an MLLM and gallery-side intra-modal cues.

3 Method

The overview of our framework is illustrated in Figure 2. Section 3.1 defines the problem setup. Section 3.2 introduces SIVA, which is trained with two complementary instructions and a three-stage curriculum to learn short-list visual associations before text-image matching. Section 3.3 presents the entire RIVA pipeline, which reorders a wide range of image candidates using long-list visual associations.

3.1 Problem Setup

We focus on the reranking problem in TIReID. Before reranking, we first train a basic VLM-based TIReID model [6, 13], $\mathcal{F}_{\text{cross}} = \{f_t, f_v\}$, on a TIReID dataset, f_t and f_v denote text and image encoders, respectively. Given a text query t and an image gallery set $\mathcal{G}$, the similarity between t and an image $v_i \in \mathcal{G}$ is computed as $s(t, v_i) = \text{Sim}(f_t(t), f_v(v_i))$, where $\text{Sim}(\cdot)$ denotes the cosine similarity. We then rank all the images in $\mathcal{G}$ in descending order of $s(t, v_j)$ and obtain the ordered top-K candidate list:

$$\mathbf{V}^K(t) = (v_{(1)}, v_{(2)}, \ldots, v_{(K)}), \quad \mathbf{s}^K(t) = (s_{(1)}, s_{(2)}, \ldots, s_{(K)}), \tag{1}$$

where $s_{(1)} \geq s_{(2)} \geq \cdots \geq s_{(K)}$.

Despite its efficiency, the discriminative power of the base TIReID model is usually limited for the fine-grained TIReID task. This means that the rank in $\mathbf{V}^K(t)$ tends to be erroneous [18].

Consequently, we propose the RIVA framework that leverages the more powerful MLLM to reorder images in $\mathbf{V}^K(t)$. However, directly feeding all K candidates into an MLLM is impractical, as its reasoning ability degrades with a large number of input images [19, 20]. To address this, RIVA adopts a local-to-global strategy. It first applies SIVA to the top-L candidates $\mathbf{V}^L(t)$ ($L < K$), where the MLLM can reliably perform visual association reasoning and identify a trustworthy anchor. Then, based on this anchor, it employs a clustering-based method to extend visual associations to the full top-K list, achieving robust reranking over $\mathbf{V}^K(t)$.

3.2 Short-List Visual Association

Directly applying a standard MLLM to TIReID reranking is suboptimal. This is mainly due to two reasons. First, TIReID relies on more fine-grained cues than a common image retrieval task, since visually similar pedestrians may differ only in local attributes, such as subtle clothing details or carried objects. Second, the reasoning ability of an MLLM tends to degrade when the number of input images is large [19, 20]. To address the above problem, we propose the **S**hort-l**I**st **V**isual **A**ssociation (**SIVA**) method that takes only a short list of L candidate images as MLLM input. We also train the SIVA model to first identify fine-grained image-image associations, then use these associations to support text-image matching.

Instruction Design. We construct two instruction-following tasks for SIVA. Task-1 is an image-image association task. Given L visually similar pedestrian images, the model is required to identify which images belong to the same identity. The instruction $\mathbf{P}_1$ is:

> *"Given the L very similar images, which images depict the same person by appearance?"*

Task-2 is a text-image matching task with image-image visual association as support. Given L images and a text description t, the model first recognizes identity-level visual associations among the images and then selects all images of the identity that best match the text query. The instruction $\mathbf{P}_2(t)$ is:

> *"Given the L very similar images, which images depict the same person by appearance? Carefully select the image(s) of the same person that more closely match the following pedestrian description: [pedestrian description]."*

Here, "[pedestrian description]" denotes t. Task-2 extends Task-1 by introducing the text query. Training in Task-1 encourages SIVA to treat identity-level image-image association as a separate visual reasoning step. Training in Task-2 further teaches SIVA to use the identified visual associations as a context for cross-modal matching.

Training Data Construction. For each paired training sample (t_i, v_i), we use t_i to rank images in $\mathcal{G}$ by similarity, giving an ordered list $\mathbf{R}_i = (v_{(1)}, \ldots, v_{(|\mathcal{G}|)})$. We randomly sample p positives and q negatives from $\mathbf{R}_i$, with $p + q = L - 1$. We then combine these sampled images with v_i to form a short image-list $\mathbf{V}_i^L$ as input for MLLM.

Finally, we denote the training data for Task-1 and Task-2 in Stage-s as

$$\mathcal{D}_1^{(s)} = \{(\mathbf{V}_i^L, \mathbf{P}_1, \mathbf{A}_{1,i})\}_{i=1}^{M_1}, \quad \mathcal{D}_2^{(s)} = \{(\mathbf{V}_i^L, \mathbf{P}_2(t_i), \mathbf{A}_{2,i})\}_{i=1}^{M_2}, \tag{2}$$

where M_1 and M_2 stand for the number of short image-lists. $\mathbf{A}_{1,i}$ and $\mathbf{A}_{2,i}$ denote the expected response of MLLM according to instruction $\mathbf{P}_1$ and $\mathbf{P}_2(t_i)$, respectively. $\mathbf{A}_{1,i}$ enumerates which images in $\mathbf{V}_i^L$ are of the same-identity. $\mathbf{A}_{2,i}$ also lists the indices of images in $\mathbf{V}_i^L$ that match t_i. $\mathbf{A}_{1,i}$ and $\mathbf{A}_{2,i}$ can be automatically obtained according to the identity labels of t_i and v_i. Examples in $\mathcal{D}_1^{(s)}$ and $\mathcal{D}_2^{(s)}$ are provided in Section B of the Appendix.

Three-Stage Training. We train the entire vision encoder and the vision-language connector of the MLLM to fully adapt it to TIReID images. We update the parameters of its LLM component using LoRA [70] to reserve LLM's generalization ability. In addition, we offer MLLM training with a three-stage curriculum. In Stage-1, we employ the large-scale synthetic HAM-PEDES dataset [17] to construct $\mathcal{D}_1^{(1)}$ and $\mathcal{D}_2^{(1)}$. In Stage-2, we construct $\mathcal{D}_1^{(2)}$ and $\mathcal{D}_2^{(2)}$ with training samples in one human-annotated TIReID database, e.g., CUHK-PEDES [9], and further finetune the model obtained in Stage-1. This step accommodates the MLLM to real-world textual descriptions and image-text modality gaps. Finally, we sample hard negatives from $\mathbf{R}_i$ in Stage-3 using the same TIReID database as in Stage-2. This step further promotes the discriminative power of the MLLM.

In each training stage, we adopt the mixture of Task-1 and Task-2 data with different ratios:

$$\text{ratio}(\mathcal{D}_1^{(s)} : \mathcal{D}_2^{(s)}) = \begin{cases} 1:1, & s = 1, \\ 1:10, & s = 2, \\ 0:1, & s = 3. \end{cases} \tag{3}$$

This curriculum first strengthens MLLM's ability in identity-level visual association between images, and then gradually shifts the learning focus toward query-conditioned text-image matching.

3.3 Long-List Visual Association

Despite SIVA's strong reasoning ability, it can only process a short list of L candidates at a time. To handle this problem, we further propose a long-list visual association method. Using visual associations between all K candidates in $\mathbf{V}^k(t)$ via clustering, it facilitates efficient reranking for a long image-list. It includes the following four steps.

Anchor Selection. An anchor image $v^\star$ is the one that is considered the most likely match in $\mathbf{V}^L(t)$ for the text query t. Following existing MLLM-based reranking methods [18], we adopt a high similarity threshold ξ to reduce unnecessary MLLM calls. Specifically, if $s_{(1)} \geq \xi$, we set $v^\star$ as $v_{(1)}$ and skip the selection of the SIVA-based anchor. Otherwise, we feed $\mathbf{V}^L(t)$ and $\mathbf{P}_2(t)$ into SIVA, and parse its response into the selected image set $\mathcal{A}(t) \subseteq \mathbf{V}^L(t)$, which is considered to contain the images that are the most likely to match t. Finally, we choose the one with the highest similarity score within $\mathcal{A}(t)$ as the anchor image:

$$v^\star = \arg\max_{v \in \mathcal{A}(t)} s(t, v). \tag{4}$$

Candidate Clustering. To efficiently exploit long-list visual associations within $\mathbf{V}^K(t)$, we cluster the visual features $\{f_v(v_{(k)})\}_{k=1}^K$ with DBSCAN [71]. We consistently set its parameter MinPts as 1, where MinPts denotes the minimum number of samples required to form a cluster. In this way, an isolated image can be retained as a singleton cluster instead of being discarded as an outlier. Finally, we denote the cluster containing the anchor image $v^\star$ as an anchor cluster $\mathcal{C}^\star$.

SIVA Checking. In this step, we carefully verify whether each image in $\mathcal{C}^\star$ matches the text query t with the powerful MLLM-based SIVA model. To achieve this, we compose short image-lists that contain L images in $\mathcal{C}^\star$. Each short image-list contains $v^\star$ by default. If there are more than L images available, we divide them into several subsets and compose multiple short image-lists. For each short image-list, we adopt SIVA to identify the image set that matches the text query t. We denote all identified images within $\mathcal{C}^\star$ as a subset $\widetilde{\mathcal{C}}^\star$.

Score Calibration. Finally, we calibrate the original image-text similarity scores as follows:

$$s'_{(k)} = s_{(k)} + \lambda \,\mathrm{conf}(\mathcal{C}^\star)\mathbb{I}(v_{(k)} \in \widetilde{\mathcal{C}}^\star), \quad 1 \leq k \leq K, \tag{5}$$

where $\lambda \in [0, 1]$ is a balancing hyperparameter and $\mathbb{I}(\cdot)$ is the indicator function. This indicates that only candidates retained in $\widetilde{\mathcal{C}}^\star$ enjoy similarity score promotion. In addition, $\mathrm{conf}(\mathcal{C}^\star) = \frac{|\widetilde{\mathcal{C}}^\star|}{|\mathcal{C}^\star|}$, which measures the consistency between DBSCAN clustering and SIVA verification. A higher $\mathrm{conf}(\mathcal{C}^\star)$ score indicates a better clustering quality and a higher probability that the images in $\mathcal{C}^\star$ match the text query t. Otherwise, the value of $\mathrm{conf}(\mathcal{C}^\star)$ would be low in order to ensure safe calibration to $s_{(k)}$.

All other images in $\mathcal{C}^\star \setminus \widetilde{\mathcal{C}}^\star$ maintain their original similarity scores. Finally, we obtain the reranked gallery image-list by sorting $\mathbf{V}^K(t)$ according to the calibrated similarity scores. In this way, we perform reranking efficiently for a long image-list via handling the sole cluster that contains the anchor image.

4 Experiments

4.1 Datasets and Metrics

Datasets. We evaluate RIVA on three standard TIReID benchmarks—CUHK-PEDES [9], ICFG-PEDES [2], and RSTPReid [21]—following their official protocols. To construct the training data for the MLLM-based SIVA model, we first adopt the large-scale synthetic HAM-PEDES [17] dataset for Stage-1 training, and then build training data for Stage-2 and -3 using the training sets of CUHK-PEDES, ICFG-PEDES, and RSTPReid, respectively.

Metrics. Following previous work [6, 13, 11, 15], we report Rank-1/5/10, mean Average Precision (mAP), and mean Inverse Negative Penalty (mINP) on all three benchmarks. Due to space limitations, detailed description of the datasets and the Rank-5 and Rank-10 results are provided in Sections A and E.1 of the Appendix.

4.2 Implementation Details

We build SIVA based on the Qwen2.5-VL-7B-Instruct model [59]. The length of the short image-list L is set to 3. We construct approximately 1M, 1.5M, and 10K short image-lists for the three training stages, respectively. It is worth noting that the data sizes of Stage-2 and -3 are computed as the total number of training samples constructed from the three benchmark datasets, whereas the MLLM is fine-tuned separately on the training set of each dataset. In Stages-1 and -2, we set the learning rates for the vision encoder, the vision-language connector, and the LLM's LoRA modules to 5×10^{-6}, 1×10^{-5} and 5×10^{-5}, respectively, with a batch size of 32. In Stage-3, we adopt slightly reduced learning rates of 4×10^{-6}, 1×10^{-5} and 4×10^{-5}, respectively. To ensure reproducibility, we employ deterministic greedy decoding [72] for the MLLM output.

We set the length of the long image-list K at 20, ξ at 0.60 and λ at 0.10. For DBSCAN [71], we adopt a fixed neighborhood radius of 0.33. Finally, to validate the effectiveness of our approach, we reproduce three state-of-the-art TIReID models, i.e. HAM [17], RDE [13], and ICL [18], and apply our RIVA reranking approach to their initial retrieval results. Following these three works, we resize all images to 384×128 pixels.

Table 1: Ablation studies on three standard benchmarks. The best scores are highlighted in bold.

Baseline	+SIVA $\mathcal{D}^{(1)}$	+SIVA $\mathcal{D}^{(2)}$	+SIVA $\mathcal{D}^{(3)}$	+Clustering	+Conf($\mathcal{C}^\star$)	CUHK-PEDES R1	CUHK-PEDES mAP	ICFG-PEDES R1	ICFG-PEDES mAP	RSTPReid R1	RSTPReid mAP
HAM [17]						77.63	69.24	67.54	41.37	71.30	55.15
HAM + Qwen2.5-VL						72.61	68.21	61.21	40.53	66.70	53.92
HAM + RIVA	✓					76.51	68.89	68.17	41.45	71.40	55.67
	✓	✓				78.30	69.43	68.91	41.87	72.25	55.89
	✓	✓	✓			79.31	70.10	69.70	42.06	73.05	56.42
	✓	✓	✓	✓		79.45	74.36	69.84	44.08	73.25	59.75
	✓	✓	✓	✓	✓	**79.84**	**76.02**	**70.36**	**45.67**	**74.05**	**61.04**

4.3 Ablation Study

In Table 1, we perform ablation studies to investigate the effectiveness of each proposed component in RIVA. The first line shows the original TIReID performance of the HAM model [17]. The second to fifth lines show the results of applying the original or our fine-tuned Qwen2.5-VL [59] model to re-rank the initial retrieval results of HAM. To achieve this, we first ask Qwen2.5-VL to identify pedestrian images within $\mathbf{V}^L(t)$ that match the text query. We then move the identified images to the top of $\mathbf{V}^L(t)$. If multiple images are identified, their ranking is based on their original similarity scores with the text query t.

It is shown that "HAM + Qwen2.5-VL" performs poorly. This is probably because general-purpose MLLMs struggle with fine-grained retrieval tasks such as TIReID. In comparison, each SIVA training stage consistently leads to a promotion in Rank-1 and mAP scores. Furthermore, our clustering-based long-list visual association method significantly promotes mAP scores by 5.92%, 3.61% and 4.62% in the three databases, respectively. The above experiments justify the effectiveness of each of our proposed components.

Then, we study the ratio of Task-1 and Task-2 data in each training stage. As shown in Figure 3, introducing more Task-1 data in the early stage improves the model's cross-modal matching ability, because it provides fine-grained supervision between visually similar images beyond the relatively coarse-grained text descriptions in TIReID. The optimal ratio gradually shifts toward Task-2 in later stages, indicating that once the model learns strong visual association ability, the training focus could move to text-image matching.

Finally, we study the optimal hyper-parameter values using the RSTPReid dataset [21]. As illustrated in Figure 4, the optimal values of ξ, λ, and K are 0.6, 0.10, and 20, respectively. Moreover, the performance of RIVA is generally robust to the change of these hyper-parameters. For space reasons, more ablation and hyper-parameter studies, e.g., the optimal value of L, are provided in Section E.2 of the Appendix.

Table 2: Comparisons with state-of-the-art TIReID methods. Image Enc. and Text Enc. denote the backbone architectures of the image and text encoders of the base TIReID model, respectively. The best results are highlighted in bold and the second-best results are underlined.

Methods	Image Enc.	Text Enc.	CUHK-PEDES			ICFG-PEDES			RSTPReid		
			R1	mAP	mINP	R1	mAP	mINP	R1	mAP	mINP
TIReID Methods without Reranking:											
Han et al. [38]	CLIP-RN	CLIP-X.	64.08	60.08	-	-	-	-	-	-	-
IVT [35]	ViT-Base	BERT	65.69	60.66	-	56.04	-	-	46.70	-	-
IRRA [6]	CLIP-ViT	CLIP-X.	73.38	66.10	50.24	63.46	38.06	7.93	60.20	47.17	25.28
CFine [39]	CLIP-ViT	BERT	69.57	-	-	60.83	-	-	50.55	-	-
CFAM [69]	CLIP-ViT	CLIP-X.	75.60	67.27	-	65.38	39.42	-	62.45	49.50	-
RDE [13]	CLIP-ViT	CLIP-X.	75.94	67.56	51.49	67.45	40.06	7.96	65.35	50.88	28.45
APTM [15]	Swin-B	BERT	76.53	66.91	-	68.51	41.22	-	67.50	52.56	-
RaSa [11]	Swin-B	BERT	76.51	69.38	-	65.28	41.29	-	66.90	52.31	-
Tan et al. [16]	CLIP-ViT	CLIP-X.	76.82	69.55	54.08	67.05	41.51	9.53	68.50	53.02	28.76
HCA [48]	CLIP-ViT	CLIP-X.	77.44	68.92	-	66.95	40.13	-	69.70	53.72	-
HAM [17]	CLIP-ViT	CLIP-X.	77.63	69.24	53.36	67.54	41.37	9.04	71.30	55.15	30.48
KPDM [41]	CLIP-ViT	CLIP-X.	78.02	69.71	53.68	69.13	42.78	9.77	67.95	52.97	30.00
DUAL [45]	CLIP-ViT	CLIP-X.	78.15	69.21	-	69.27	42.56	-	70.40	54.83	-
ICL [18]	CLIP-ViT	CLIP-X.	78.17	69.71	53.55	69.50	42.32	8.96	70.10	53.65	29.59
TIReID Methods with Reranking:											
ICL + THI [18]	CLIP-ViT	CLIP-X.	79.06	70.58	54.79	70.07	42.96	9.20	72.15	54.41	30.21
HAM [17] + RIVA	CLIP-ViT	CLIP-X.	79.84	**76.02**	**64.69**	70.36	45.67	**10.21**	**74.05**	**61.04**	**36.49**
ICL [18] + RIVA	CLIP-ViT	CLIP-X.	**79.86**	74.87	63.23	**71.32**	**45.96**	10.07	72.60	60.47	36.44

Table 3: More comparisons between RIVA and THI with different base TIReID models.

Baseline	Rerank methods	CUHK-PEDES		ICFG-PEDES		RSTPReid		ΔAvg
		R1	mAP	R1	mAP	R1	mAP	
IRRA [6]	-	73.38	66.10	63.46	38.06	60.20	47.17	-
	+THI [18]	76.07	67.39	65.19	38.24	63.85	48.56	+1.82
	+RIVA	**76.82**	**73.27**	**66.56**	**42.09**	**65.15**	**56.18**	**+5.28**
RDE [13]	-	75.94	67.56	67.45	40.06	65.35	50.88	-
	+THI [18]	77.40	68.70	68.89	40.90	68.30	53.10	+1.68
	+RIVA	**78.22**	**74.63**	**69.92**	**44.73**	**69.25**	**59.24**	**+4.79**
HAM [17]	-	77.63	69.24	67.54	41.37	71.30	55.15	-
	+THI [18]	78.46	69.94	68.35	41.91	73.30	55.85	+0.93
	+RIVA	**79.84**	**76.02**	**70.36**	**45.67**	**74.05**	**61.04**	**+4.13**

4.4 Comparisons with State-of-the-Art Methods

Comparisons in traditional evaluation settings. We apply our RIVA reranking approach to the initial retrieval results by HAM [17] and ICL [18], respectively. As shown in Table 2, RIVA consistently improves Rank-1, mAP, and mINP scores in all three benchmarks, leading to the best overall performance in each data set compared to previous works. In particular, when applied to HAM [17], RIVA increases its Rank-1 accuracy by 2.21%, 2.82% and 2.75%, mAP by 6.78%, 4.30% and 5.89%, and mINP by 11.33%, 1.17% and 6.01%, respectively. Similar gains are observed in ICL. Compared to the existing THI reranking approach [18], RIVA consistently delivers larger performance gains across all datasets and metrics. In particular, "ICL+RIVA" outperforms "ICL+THI" by 4.29%,

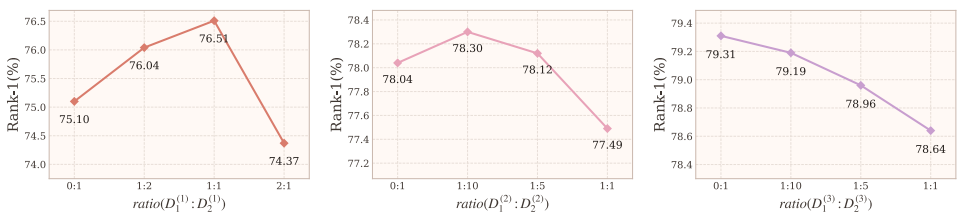

Figure 3: Study on the ratio between Task-1 and Task-2 data in each training stage of SIVA.

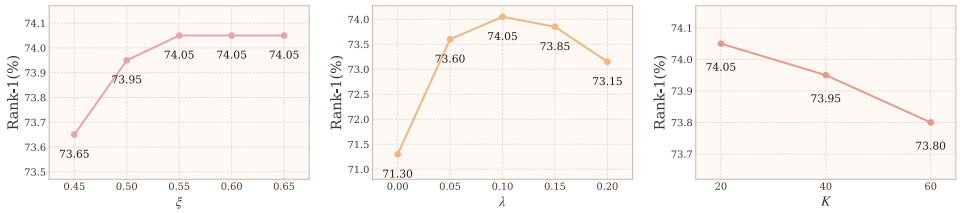

Figure 4: Hyper-parameter study on the RSTPReid dataset [21].

Table 4: Comparisons of generalization capabilities among three standard benchmarks.

Methods	Training Sets	CUHK-PEDES		ICFG-PEDES		RSTPReid	
		R1	mAP	R1	mAP	R1	mAP
IRRA [6]	CUHK-PEDES	73.38	66.10	42.41	21.77	53.25	39.63
	ICFG-PEDES	33.48	31.56	63.46	38.06	45.30	36.82
	RSTPReid	32.80	30.29	32.30	20.54	60.20	47.17
ICL + THI [18]	CUHK-PEDES	77.91	69.13	52.80	25.60	61.30	43.42
	ICFG-PEDES	49.29	40.82	69.02	41.21	60.15	43.72
	RSTPReid	47.35	38.91	50.52	29.26	70.55	53.68
HAM [17]	CUHK-PEDES	77.63	69.23	59.67	34.21	60.25	46.23
	ICFG-PEDES	64.02	57.65	67.54	41.37	58.10	44.33
	RSTPReid	64.95	58.80	56.04	35.45	71.30	55.15
HAM [17]+RIVA	CUHK-PEDES	**79.84**	**76.02**	**62.64**	**36.24**	**64.50**	**50.89**
	ICFG-PEDES	**67.40**	**63.68**	**70.36**	**45.67**	**63.35**	**52.06**
	RSTPReid	**67.46**	**62.99**	**59.05**	**38.13**	**74.05**	**61.04**

3.00% and 6.06% in mAP on the three databases, respectively. These results indicate that exploiting visual associations between candidate images is essential to reduce the image-text modality gap and contributes significantly to effective reranking.

We perform more comparisons between RIVA and THI [18] with three representative TIReID base models: IRRA [6], RDE [13], and HAM [17]. We summarize the experimental results in Table 3. RIVA is shown to consistently yield larger performance gains than THI in all three base models. In particular, it improves the performance of IRRA and RDE by 5.28% and 4.79% on average, surpassing THI's 1.82% and 1.68%; It also improves HAM's performance by 4.13% on average, significantly outperforming THI's 0.93%. The above experimental results further justify the effectiveness of RIVA.

Comparisons in cross-dataset evaluation settings. In Table 4, we conduct cross-dataset experiments to compare the generalizability of different TIReID methods. RIVA is shown to consistently achieve the highest Rank-1 and mAP scores. For example, when we perform training on ICFG-PEDES and evaluation on RSTPReid, RIVA improves HAM's Rank-1 and mAP scores by 5.25% and 7.73%, respectively. These results highlight the strong generalizability of RIVA.

4.5 Inference Efficiency

We compare RIVA with THI [18] under matched hardware, image-resolution, and decoding settings. Table 5 reports wall time and query throughput as mean $\pm$ standard deviation over three runs on 2$\times$ GeForce RTX 3090 GPUs, using tensor parallelism of 2 and BF16 precision. Calls are averaged per query.

Table 5: Inference efficiency of THI and RIVA. Time is the wall time of the evaluated query workload; throughput is measured in queries per second. Calls/q denotes the average number of MLLM requests per query.

Dataset	Method	Time (s)	Queries/s	Calls/q
CUHK-PEDES	THI	1223.20$\pm$18.07	5.033$\pm$0.074	4.241
	RIVA	**846.35$\pm$0.67**	**7.274$\pm$0.006**	**2.194**
UFine6926	THI	3402.60$\pm$20.81	4.484$\pm$0.027	3.564
	RIVA	**2880.95$\pm$14.81**	**5.296$\pm$0.027**	**2.635**

RIVA achieves 1.445× and 1.181× the throughput of THI on CUHK-PEDES and UFine6926, respectively, while reducing wall time by 30.8% and 15.3%. It requires 48.3% and 26.1% fewer MLLM calls per query, respectively. This reduction follows from using short-list anchor selection and verifying only the anchor cluster, with confident queries bypassing anchor selection. Gallery image features can be precomputed, and query-dependent clustering is restricted to the top-K candidates. Together with the retrieval improvements in Table 2 and Section E.7, these results support the use of RIVA in server-side retrieval applications that prioritize accuracy at moderate query loads. Additional controls and deployment assumptions are discussed in Sections E and F.3.

5 Conclusion

In this paper, we identify a key limitation of existing TIReID methods as their strict instance-wise matching paradigm, which overlooks intra-modal visual associations among top-ranked gallery candidates. To address this, we propose RIVA, a plug-and-play reranking framework that leverages such associations to mitigate the inter-modal semantic gap. RIVA improves reranking through two complementary components: SIVA, an MLLM-based short-list reranker trained with a dual-task, three-stage curriculum that learns identity-level image-image associations as visual context for fine-grained text-image matching, and a clustering-based long-list visual association module that efficiently extends reranking to a wider range of candidates via confidence-guided score calibration. We validate RIVA's effectiveness through extensive experiments on three standard benchmarks. RIVA consistently improves multiple base TIReID models and outperforms existing reranking methods in both traditional and cross-dataset settings. **Limitations.** As an MLLM-based reranking framework, RIVA introduces additional inference latency, and we leave exploration of more efficient inference strategies for future work. **Broader Impacts.** RIVA improves real-world text-based person retrieval systems, benefiting applications such as video surveillance and public security. We encourage the responsible use of this technique. To the best of our knowledge, this work has no obvious negative social impacts.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62476099 and 62076101, Guangdong Basic and Applied Basic Research Foundation under Grant 2024B1515020082, the TCL Young Scholars Program, and the High Performance Computing Platform of South China University of Technology. The authors have no competing interests to declare.

References

[1] Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. VITAA: Visual-textual attributes alignment in person search by natural language. In *ECCV*, 2020.

[2] 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.

[3] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *ACM MM*, 2022.

[4] Ammarah Farooq, Muhammad Awais, Josef Kittler, and Syed Safwan Khalid. AXM-Net: Implicit cross-modal feature alignment for person re-identification. In *AAAI*, 2022.

[5] Shuanglin Yan, Neng Dong, Jun Liu, Liyan Zhang, and Jinhui Tang. Learning comprehensive representations with richer self for text-to-image person re-identification. In *ACM MM*, pages 6202–6211, 2023.

[6] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *CVPR*, 2023.

[7] 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 *ICCV*, 2023.

[8] Maryam Bukhari, Sadaf Yasmin, Sheneela Naz, Muazzam Maqsood, Jehyeok Rew, and Seungmin Rho. Language and vision based person re-identification for surveillance systems using deep learning with LIP layers. *Image and Vision Computing*, 132:104658, 2023.

[9] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, 2017.

[10] Hiren Galiyawala and Mehul S. Raval. Person retrieval in surveillance using textual query: A review. *Multimedia Tools and Applications*, 2021.

[11] 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 *IJCAI*, 2023.

[12] Dixuan Lin, Yi-Xing Peng, Jingke Meng, and Wei-Shi Zheng. Cross-modal adaptive dual association for text-to-image person retrieval. *IEEE TMM*, 26:6609–6620, 2024.

[13] 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 *CVPR*, pages 27197–27206, 2024.

[14] Xinpan Yuan, Shaomin Xie, Liujie Hua, Chengyuan Zhang, Guihu Zhao, and Lin Yuanbo Wu. Geometry-aware noisy correspondence mitigation for cross-modal text-based person retrieval. In *AAAI*, volume 40, pages 12268–12276, 2026.

[15] 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 *ACM MM*, 2023.

[16] Wentao 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 *CVPR*, pages 17127–17137, 2024.

[17] 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 *CVPR*, pages 9220–9230, 2025.

[18] 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 *CVPR*, pages 14390–14399, 2025.

[19] Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. MileBench: Benchmarking MLLMs in long context. *arXiv preprint arXiv:2404.18532*, 2024.

[20] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. In *ICLR*, 2025.

[21] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. DSSL: Deep surroundings-person separation learning for text-based person retrieval. In *ACM MM*, 2021.

[22] Yuyu Wang, Chunjuan Bo, Dong Wang, Shuang Wang, Yunwei Qi, and Huchuan Lu. Language person search with mutually connected classification loss. In *ICASSP*, 2019.

[23] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM TOMM*, 2020.

[24] Nikolaos Sarafianos, Xiang Xu, and Ioannis A. Kakadiaris. Adversarial representation learning for text-to-image matching. In *ICCV*, 2019.

[25] Tianlang Chen, Chenliang Xu, and Jiebo Luo. Improving text-based person search by spatial matching and adaptive threshold. In *WACV*, 2018.

[26] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *ECCV*, 2018.

[27] Shiping Li, Min Cao, and Min Zhang. Learning semantic-aligned feature representation for text-based person search. In *ICASSP*, 2022.

[28] Kai Niu, Linjiang Huang, Yan Huang, Peng Wang, Liang Wang, and Yanning Zhang. Cross-modal co-occurrence attributes alignments for person search by language. In *ACM MM*, pages 4426–4434, 2022.

[29] Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. Pose-guided multi-granularity attention network for text-based person search. In *AAAI*, 2020.

[30] Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. TIPCB: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 2022.

[31] Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. LapsCore: Language-guided person search via color reasoning. In *ICCV*, 2021.

[32] Wei Suo, Mengyang Sun, Kai Niu, Yiqi Gao, Peng Wang, Yanning Zhang, and Qi Wu. A simple and robust correlation filtering method for text-based person search. In *ECCV*, 2022.

[33] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. CAIBC: Capturing all-round information beyond color for text-based person retrieval. In *ACM MM*, 2022.

[34] Jicheol Park, Dongwon Kim, Boseung Jeong, and Suha Kwak. PLOT: Text-based person search with part slot attention for corresponding part discovery. In *ECCV*, pages 474–490, 2024.

[35] Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. See finer, see more: Implicit modality alignment for text-based person retrieval. In *ECCVW*, 2023.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[37] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven C. H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

[38] Xiao Han, Sen He, Li Zhang, and Tao Xiang. Text-based person search with limited data. *arXiv preprint arXiv:2110.10807*, 2021.

[39] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. CLIP-driven fine-grained text-image person re-identification. *IEEE TIP*, 32:6032–6046, 2023.

[40] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of CLIP for text-based person search. In *AAAI*, pages 465–473, 2024.

[41] Shaofeng You, Tianle Miao, Qihang Chen, Xin Li, Zhuo Cheng, and Dapeng Luo. KPDM: Key phrase dynamic masking for robust text-to-image person retrieval. In *AAAI*, volume 40, pages 12099–12107, 2026.

[42] Shenshen Li, Xing Xu, Yang Yang, Fumin Shen, Yijun Mo, Yujie Li, and Heng Tao Shen. DCEL: Deep cross-modal evidential learning for text-based person retrieval. In *ACM MM*, pages 6292–6300, 2023.

[43] Shenshen Li, Chen He, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. Adaptive uncertainty-based learning for text-based person retrieval. In *AAAI*, pages 3172–3180, 2024.

[44] Zhiwei Zhao, Bin Liu, Yan Lu, Qi Chu, and Nenghai Yu. Unifying multi-modal uncertainty modeling and semantic alignment for text-to-image person re-identification. In *AAAI*, volume 38, pages 7534–7542, 2024.

[45] Yufei Zheng, Jiawei Liu, Bingyu Hu, Zikun Wei, Yong Wu, and Zheng-Jun Zha. Dual uncertainty-guided feature alignment learning for text-based person retrieval. In *ACM MM*, pages 6037–6046, 2025.

[46] Shuanglin Yan, Jun Liu, Neng Dong, Liyan Zhang, and Jinhui Tang. Prototypical prompting for text-to-image person re-identification. In *ACM MM*, pages 2331–2340, 2024.

[47] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Wei-Shi Zheng, and Nong Sang. Weakly supervised text-based person re-identification. In *ICCV*, pages 11395–11404, 2021.

[48] Yifei Deng, Chenglong Li, Futian Wang, and Jin Tang. Learning hierarchical cross-modal association with intra-modal context for text-image person retrieval. In *ACM MM*, pages 2723–2731, 2025.

[49] Yiwei Ma, Xiaoshuai Sun, Jiayi Ji, Guannan Jiang, Weilin Zhuang, and Rongrong Ji. BEAT: Bi-directional one-to-many embedding alignment for text-based person retrieval. In *ACM MM*, pages 4157–4168, 2023.

[50] Jialong Zuo, Changqian Yu, Nong Sang, and Changxin Gao. PLIP: Language-image pre-training for person representation learning. *arXiv preprint arXiv:2305.08386*, 2023.

[51] Yang Bai, Jingyao Wang, Min Cao, Chen Chen, Ziqiang Cao, Liqiang Nie, and Min Zhang. Text-based person search without parallel image-text data. In *ACM MM*, pages 757–767, 2023.

[52] Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, and Mang Ye. Chat-based person retrieval via dialogue-refined cross-modal alignment. In *CVPR*, pages 3952–3962, 2025.

[53] Yiding Lu, Mouxing Yang, Dezhong Peng, Peng Hu, Yijie Lin, and Xi Peng. LLaVA-ReID: Selective multi-image questioner for interactive person re-identification. *arXiv preprint arXiv:2504.10174*, 2025.

[54] Hongyu Liu, Hongwei Ge, Yuxuan Liu, and Yaqing Hou. Dialogue-driven interactive dynamic learning for text-to-image person retrieval. In *ACM MM*, pages 9177–9185, 2025.

[55] Jintao Sun, Hao Fei, Gangyi Ding, and Zhedong Zheng. From data deluge to data curation: A Filtering-WoRA paradigm for efficient text-based person search. In *WWW*, pages 2341–2351, 2025.

[56] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022.

[57] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[58] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[59] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[60] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

[61] Kimi Team. Kimi K2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

[62] GLM-V Team. GLM-5V-Turbo: Toward a native foundation model for multimodal agents. *arXiv preprint arXiv:2604.26752*, 2026.

[63] Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. MLLM is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*, 2024.

[64] Antoine Chaffin and Aurélien Lac. MonoQwen: Visual document reranking. LightOn, 2024.

[65] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. MM-Embed: Universal multimodal retrieval with multimodal LLMs. *arXiv preprint arXiv:2411.02571*, 2024.

[66] Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. LAMRA: Large multimodal model as your advanced retrieval assistant. In *CVPR*, pages 4015–4025, 2025.

[67] Mingjun Xu, Jinhan Dong, Jue Hou, Zehui Wang, Sihang Li, Zhifeng Gao, Renxin Zhong, and Hengxing Cai. MM-R5: Multimodal reasoning-enhanced reranker via reinforcement learning for document retrieval. *arXiv preprint arXiv:2506.12364*, 2025.

[68] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

[69] 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 *CVPR*, pages 22010–22019, 2024.

[70] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[71] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996.

[72] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020.

A Dataset Descriptions

CUHK-PEDES. CUHK-PEDES [9] serves as a pioneering benchmark for TIReID. It consists of 40,206 images that span 13,003 identities in total, each image labeled with two textual descriptions. Following the official split, the training set contains 34,054 images (11,003 identities), the validation set 3,078 images (1,000 identities), and the test set 3,074 images (1,000 identities).

ICFG-PEDES. ICFG-PEDES [2] comprises 54,522 images covering 4,102 unique identities, each image annotated by a detailed textual description. It is split into a training subset of 34,674 image-text pairs for 3,102 identities and a test subset of 19,848 pairs for the remaining 1,000 identities.

RSTPReid. RSTPReid [21] is a multimodal person re-identification dataset with 20,505 images taken by 15 cameras, covering 4,101 unique identities. Each identity is represented by five cross-camera images, and each image is paired with two textual captions. Under the official train-val-test partition, the training set contains 3,701 identities, while the validation and test sets each include 200 identities.

HAM-PEDES. HAM-PEDES [17] is a large-scale synthetic TIReID dataset built for model pre-training purposes. It contains 1.0 million pedestrian images sampled from SYNTH-PEDES [50], each associated with two stylistically diverse captions. These descriptions are generated by LLaVA-1.6 that mimics the annotation preferences of thousands of human annotators, using 3,000 style prompts to ensure rich and diverse textual expressions.

B Training Set Examples

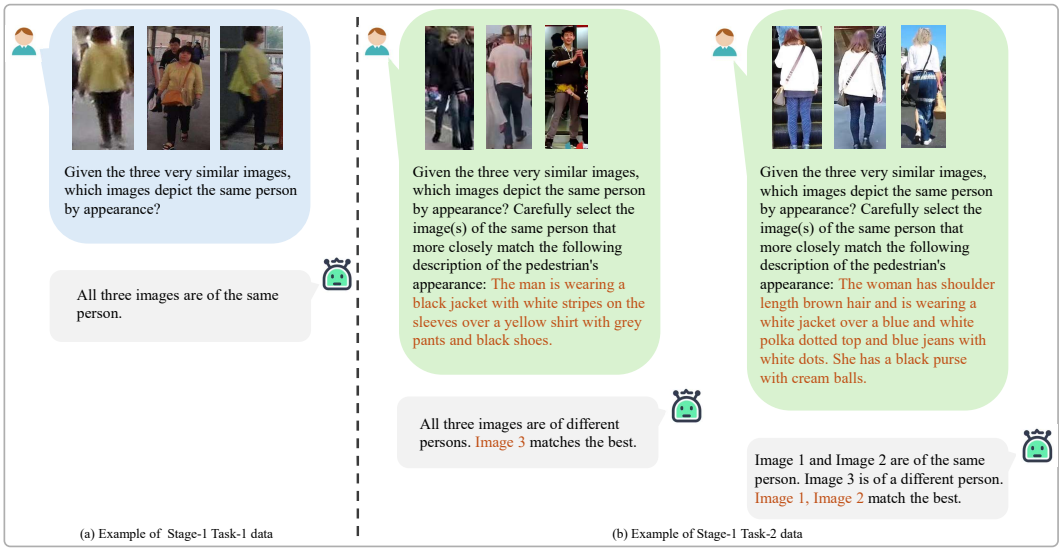

Figure 5: Examples of Stage-1 data when the value of L is 3.

In this section, we present representative examples from the SIVA training dataset. As shown in Figure 5(a) and Figure 6(a), Task-1 focuses on recognition of intra-modal identities within the visual modality. Task-2 leverages the visual association learned by Task-1 as a context to facilitate the inter-modal alignment. As illustrated in Figure 5, short input image-lists can be categorized into three types according to the number of matched images they contain: one, two, or three matched images. The model is required to identify all images of the pedestrian that match the textual description.

In Figure 6, we show examples of Stage-2 Task-1 data, Stage-2 Task-2 data, and Stage-3 Task-2 data constructed from the same textual query and ground-truth pedestrian. In Figure 6(a), Task-1 is essentially ambiguous under the textual condition alone, since multiple candidates may satisfy the coarse-grained textual description. However, the images provide richer fine-grained visual details,

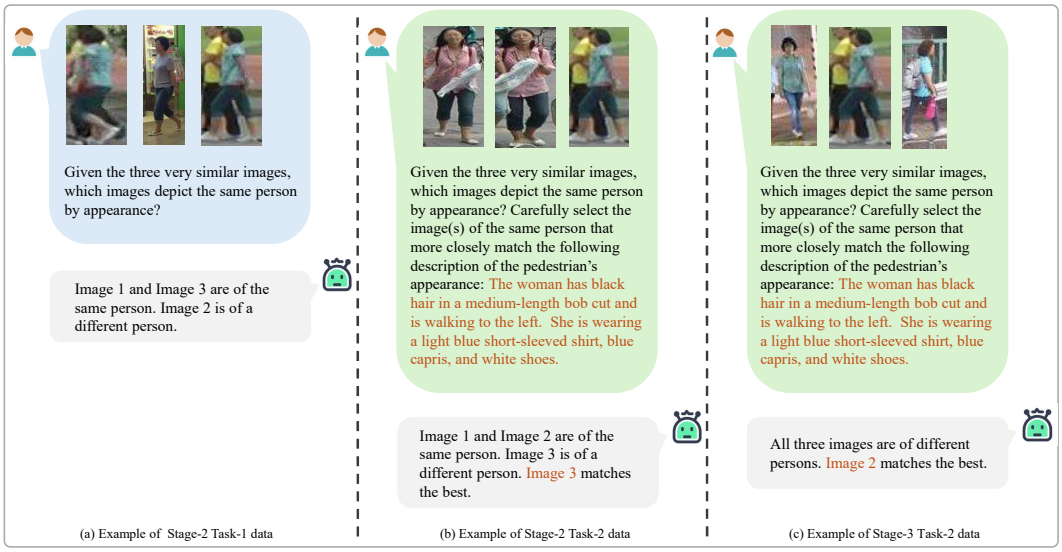

Figure 6: Examples of Stage-2 and Stage-3 data when the value of L is 3.

which make it possible to distinguish different identities. Therefore, we construct a more challenging unimodal task to mitigate the limitations of the coarse text modality, which may otherwise introduce false negatives, and to encourage the model to learn more robust visual representations, as illustrated in Figure 2(a). As shown in Figure 6(b) and (c), Stage-3 is more challenging than Stage-2 under the same textual query. Since the clothing colors of the images in Figure 6(c) are highly similar, the model has to leverage its spatial perception ability to determine that the person in the second image is walking toward the left side of the image. We also vary the number of images per identity and allow multiple negatives from the same identity to co-occur, preventing the model from using shortcuts such as selecting the only repeated identity, as illustrated in Figure 2(b).

Through this multi-task, multi-stage curriculum learning, the model ultimately achieves strong performance in TIReID.

C Reranking Result Examples

In this section, we present realistic reranking examples of RIVA on the test set of CUHK-PEDES. As illustrated in Figure 7, given a pedestrian text query, we first perform an initial retrieval using a VLM-based TIReID model, such as IRRA [6] or RDE [13]. We then employ SIVA to select an anchor image according to Eq. 4. Based on the visual features extracted by the VLM image encoder, we further perform clustering over the visual features of top-K retrieved candidate images. After estimation of the cluster confidence and refinement with SIVA, we obtain the final clustering results $\widetilde{\mathcal{C}}^{\star}$ and the confidence score $\mathrm{conf}(\mathcal{C}^{\star})$ for the anchor cluster. The confidence scores are shown as the numbers in the green boxes. Finally, based on the selected anchor cluster and confidence score, we update the similarity scores by Eq. 5 and obtain the final ranking.

As shown in Figure 8, cluster confidence helps mitigate the negative effect caused by an incorrectly selected anchor image from SIVA. When SIVA selects a wrong anchor, we can still use other images within the same cluster to assess whether the identity associated with that anchor is consistent with the text description. Moreover, through cluster refinement, we remove the non-anchor images from this incorrect cluster. Even when the original DBSCAN clustering is accurate, this refinement step prevents other images in the wrong cluster from being promoted in the ranking, thereby reducing contamination of the top-ranked results. For example, as illustrated in Figure 8, SIVA compares the other images in the anchor cluster with the text query, removes the non-anchor images from the incorrect cluster, and assigns it a low confidence score of 0.25, thereby preventing the erroneous images from being propagated toward higher ranks. Figure 9 further shows that when DBSCAN produces inaccurate clustering results, cluster refinement can correct the cluster structure by removing images from other clusters. This leads to more accurate reranking and better final retrieval results.

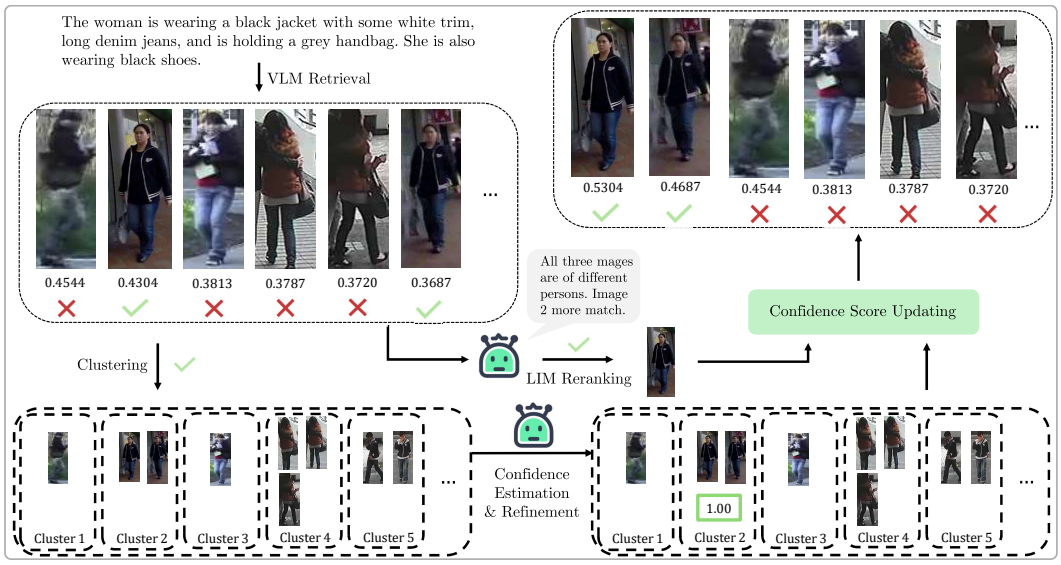

Figure 7: Example of the proposed reranking process. Given a pedestrian text query, we first conduct VLM retrieval and cluster top-K candidates. SIVA is used for anchor selection and cluster confidence estimation and refinement. For efficiency, confidence can be computed only for the anchor-associated cluster. Final rankings are obtained via similarity score updating.

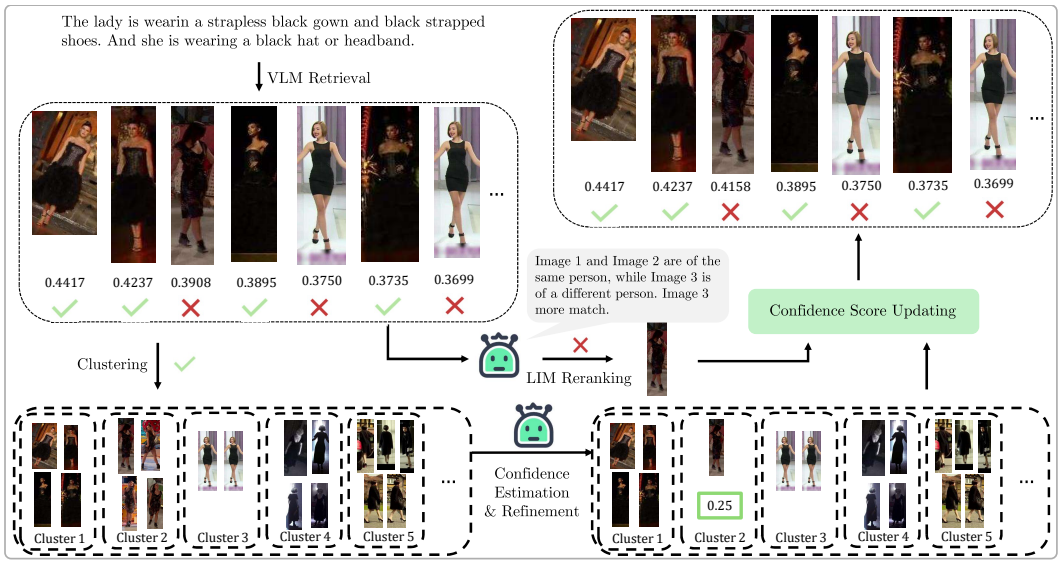

Figure 8: Example of the proposed reranking process when SIVA select the wrong anchor.

D Additional Implementation Details

For SIVA's fine-tuning, the LLM component is updated via LoRA [70] with rank $r = 64$ and scaling factor $\alpha = 128$. Training proceeds in three stages, each running for 2 epochs. In Stages-1 and -2, MLLM is fine-tuned with an effective batch size of 32. Stage-3 uses an effective batch size 16 to stabilize refinement. All stages employ a warmup ratio of 0.05. The default global seed is 42. Specifically, Stage-1 contains a total of 1,063,218 short image lists. Stage-2 contains 749,276, 381,414, and 407,110 short image lists constructed from CUHK-PEDES, ICFG-PEDES, and RSTPReid, respectively, while Stage 3 contains 3,625, 3,616, and 3,543 short image lists for the three datasets, respectively.

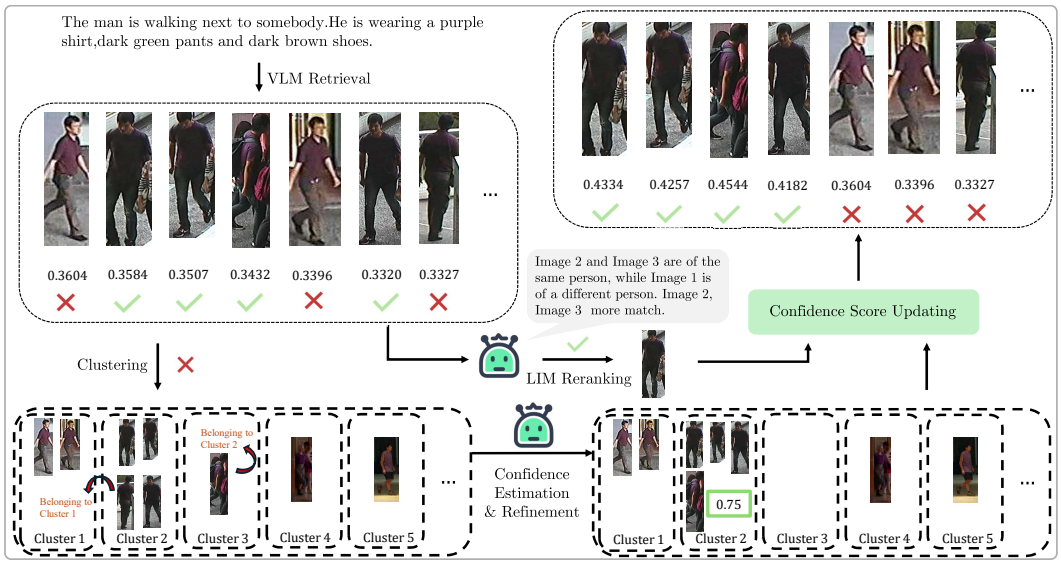

Figure 9: Example of the proposed reranking process under clustering errors.

Besides, we follow the official configuration with IRRA for HAM [17]. For IRRA [6], the [CLS] token is used as the visual representation for clustering. For RDE [13], BGE features are adopted as the visual representation. When reproducing ICL [18], we follow the official setting and initialize its VLM with NAM [16]. SIVA training uses 8× GeForce RTX 3090 GPUs. The inference-cost comparison in Section 4.5 uses 2× RTX 3090 GPUs.

E Additional Experimental Results

E.1 More Detailed Experimental Results

We report more detailed results of RIVA in the CUHK-PEDES, ICFG-PEDES, and RSTPReid databases in Table 6, Table 7, and Table 8, respectively. As shown in these tables, RIVA consistently improves Rank-1, Rank-10, mAP, and mINP scores over the base TIReID models on all three benchmarks. We note that there is slight degradation in Rank-5 accuracy. Our further analysis shows that this is often caused by ambiguous or noisy textual descriptions. In these cases, it is challenging or even impossible for RIVA to identify an anchor image consistent with the ground-truth, which degrades the effect of score calibration and finally leads to a slight decrease in Rank-5 accuracy. Indeed, the same phenomenon happens to the THI reranking approach [18] as well.

E.2 Parameter Study on Short Image-List Length

In this experiment, we investigate the optimal value of L. We adopt the RSTPReid database to construct the Stage-2 training data. We employ these training data alone to train our SIVA model, and perform evaluation on the RSTPReid testing set. Figure 10 shows that the optimal value of L is 3. This result indicates that the reasoning ability of MLLMs degrades with a large number of input images. This phenomenon is consistent with the discoveries in [19, 20].

E.3 Comparison with Feature-Space Reranking

We compare RIVA with training-free reranking methods using the same HAM retrieval results, top-K lists, and visual features. The hyperparameters of these controls are selected on validation data. Average query expansion aggregates candidate features to refine retrieval, k-reciprocal reranking uses reciprocal neighborhoods, and graph diffusion propagates feature-space similarity. We also include a DBSCAN-only control that uses the original rank-1 image as the anchor and uniformly promotes its cluster without any MLLM calls, with the same clustering parameters and score-calibration strength as RIVA.

Table 6: Comparisons with state-of-the-art TIReID methods on the CUHK-PEDES database.

Methods	Venue	Image Enc.	Text Enc.	R1	R5	R10	mAP	mINP
TIReID Methods without Reranking:								
Han et al. [38]	arXiv'21	CLIP-RN	CLIP-X.	64.08	81.73	88.19	60.08	-
IVT [35]	ECCV'22	ViT-Base	BERT	65.69	85.93	91.15	60.66	-
IRRA [6]	CVPR'23	CLIP-ViT	CLIP-X.	73.38	89.93	93.71	66.10	50.24
CFine [39]	TIP'23	CLIP-ViT	BERT	69.57	85.93	91.15	-	-
CFAM [69]	CVPR'24	CLIP-ViT	CLIP-X.	75.60	90.53	94.36	67.27	-
RDE [13]	CVPR'24	CLIP-ViT	CLIP-X.	75.94	90.14	94.12	67.56	51.49
APTM [15]	ACM MM'23	Swin-B	BERT	76.53	90.04	94.15	66.91	-
RaSa [11]	IJCAI'23	Swin-B	BERT	76.51	90.29	94.25	69.38	-
Tan et al. [16]	CVPR'24	CLIP-ViT	CLIP-X.	76.82	91.16	94.46	69.55	54.08
HCA [48]	ACM MM'25	CLIP-ViT	CLIP-X.	77.44	91.23	94.61	68.92	-
HAM [17]	CVPR'25	CLIP-ViT	CLIP-X.	77.63	91.57	94.82	69.24	53.36
KPDM [41]	AAAI'26	CLIP-ViT	CLIP-X.	78.02	91.49	95.00	69.71	53.68
DUAL [45]	ACM MM'25	CLIP-ViT	CLIP-X.	78.15	91.18	94.54	69.21	-
ICL [18]	CVPR'25	CLIP-ViT	CLIP-X.	78.17	91.47	94.93	69.71	53.55
TIReID Methods with Reranking:								
ICL + THI [18]	CVPR'25	CLIP-ViT	CLIP-X.	79.06	91.33	94.79	70.58	54.79
HAM [17] + RIVA	Ours	CLIP-ViT	CLIP-X.	79.84	91.15	94.90	76.02	64.69
ICL [18] + RIVA	Ours	CLIP-ViT	CLIP-X.	79.86	91.52	95.14	74.87	63.23

Table 7: Comparisons with state-of-the-art TIReID methods on the ICFG-PEDES database.

Methods	Venue	Image Enc.	Text Enc.	R1	R5	R10	mAP	mINP
TIReID Methods without Reranking:								
IVT [35]	ECCV'22	ViT-Base	BERT	56.04	73.60	80.22	-	-
IRRA [6]	CVPR'23	CLIP-ViT	CLIP-X.	63.46	80.25	85.82	38.06	7.93
CFine [39]	TIP'23	CLIP-ViT	BERT	60.83	76.55	82.42	-	-
CFAM [69]	CVPR'24	CLIP-ViT	CLIP-X.	65.38	81.17	86.35	39.42	-
RDE [13]	CVPR'24	CLIP-ViT	CLIP-X.	67.45	82.47	87.36	40.06	7.96
APTM [15]	ACM MM'23	Swin-B	BERT	68.51	82.99	87.56	41.22	-
RaSa [11]	IJCAI'23	Swin-B	BERT	65.28	80.40	85.12	41.29	-
Tan et al. [16]	CVPR'24	CLIP-ViT	CLIP-X.	67.05	82.16	87.33	41.51	9.53
HCA [48]	ACM MM'25	CLIP-ViT	CLIP-X.	66.95	82.34	87.06	40.13	-
HAM [17]	CVPR'25	CLIP-ViT	CLIP-X.	67.54	82.98	87.74	41.37	9.04
KPDM [41]	AAAI'26	CLIP-ViT	CLIP-X.	69.13	83.46	88.10	42.78	9.77
DUAL [45]	ACM MM'25	CLIP-ViT	CLIP-X.	69.27	83.00	87.56	42.56	-
ICL [18]	CVPR'25	CLIP-ViT	CLIP-X.	69.50	83.40	88.01	42.32	8.96
TIReID Methods with Reranking:								
ICL + THI [18]	CVPR'25	CLIP-ViT	CLIP-X.	70.07	83.16	87.86	42.96	9.20
HAM [17] + RIVA	Ours	CLIP-ViT	CLIP-X.	70.36	82.85	87.96	45.67	10.21
ICL [18] + RIVA	Ours	CLIP-ViT	CLIP-X.	71.32	82.91	88.04	45.96	10.07

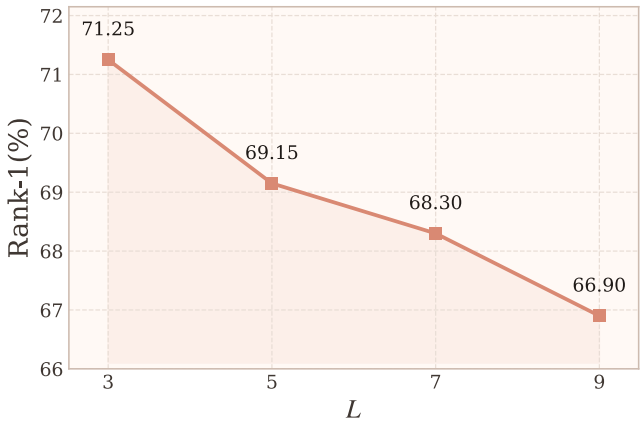

Figure 10: Parameter study on short image-list length L.

Table 8: Comparisons with state-of-the-art TIReID methods on the RSTPReid database.

Methods	Venue	Image Enc.	Text Enc.	R1	R5	R10	mAP	mINP
TIReID Methods without Reranking:								
IVT [35]	ECCV'22	ViT-Base	BERT	46.70	70.00	78.80	-	-
IRRA [6]	CVPR'23	CLIP-ViT	CLIP-X.	60.20	81.35	88.05	47.17	25.28
CFine [39]	TIP'23	CLIP-ViT	BERT	50.55	72.50	81.60	-	-
CFAM [69]	CVPR'24	CLIP-ViT	CLIP-X.	62.45	83.55	91.10	49.50	-
RDE [13]	CVPR'24	CLIP-ViT	CLIP-X.	65.35	83.95	89.90	50.88	28.45
APTM [15]	ACM MM'23	Swin-B	BERT	67.50	85.70	91.45	52.56	-
RaSa [11]	IJCAI'23	Swin-B	BERT	66.90	86.50	91.35	52.31	-
Tan et al. [16]	CVPR'24	CLIP-ViT	CLIP-X.	68.50	87.15	92.10	53.02	28.76
HCA [48]	ACM MM'25	CLIP-ViT	CLIP-X.	69.70	86.95	92.45	53.72	-
HAM [17]	CVPR'25	CLIP-ViT	CLIP-X.	71.30	87.05	92.50	55.15	30.48
KPDM [41]	AAAI'26	CLIP-ViT	CLIP-X.	67.95	86.35	91.65	52.97	30.00
DUAL [45]	ACM MM'25	CLIP-ViT	CLIP-X.	70.40	87.65	92.50	54.83	-
ICL [18]	CVPR'25	CLIP-ViT	CLIP-X.	70.10	85.95	91.15	53.65	29.59
TIReID Methods with Reranking:								
ICL + THI [18]	CVPR'25	CLIP-ViT	CLIP-X.	72.15	85.85	90.85	54.41	30.21
HAM [17] + RIVA	Ours	CLIP-ViT	CLIP-X.	74.05	87.25	92.75	61.04	36.49
ICL [18] + RIVA	Ours	CLIP-ViT	CLIP-X.	72.60	86.05	91.50	60.47	36.44

Table 9: Comparison with training-free reranking on HAM retrieval results. Each entry reports R1 / mAP (%). The best results are highlighted in bold.

Reranker	CUHK-PEDES	ICFG-PEDES	RSTPReid
None	77.63 / 69.24	67.54 / 41.37	71.30 / 55.15
Average query expansion	77.63 / 73.25	67.54 / 43.86	71.30 / 59.66
k-reciprocal	74.45 / 71.73	63.59 / 41.67	67.50 / 58.84
Graph diffusion	61.19 / 59.23	64.80 / 42.81	54.70 / 47.95
DBSCAN only	77.63 / 72.44	67.54 / 43.10	71.30 / 58.13
SIVA only	79.31 / 70.10	69.70 / 42.06	73.05 / 56.42
RIVA	**79.84 / 76.02**	**70.36 / 45.67**	**74.05 / 61.04**

As shown in Table 9, feature-space reranking mainly improves mAP, while SIVA yields larger Rank-1 gains. In particular, DBSCAN-only promotion improves mAP but leaves Rank-1 unchanged. It does not provide query-conditioned anchor correction or member verification. RIVA combines these operations with long-list association and achieves the highest Rank-1 and mAP across the three datasets. These results complement the component ablation in Table 1.

E.4 Long-List Neighborhood Construction

To examine the choice of long-list association strategy, we keep the SIVA model fixed and replace only the construction of the anchor-conditioned candidate neighborhood. All variants use the same HAM retrieval results on CUHK-PEDES, top-K lists, CLIP-ViT image features, SIVA checkpoint, and selected anchor image. The resulting neighborhoods undergo the same SIVA-based member verification and score calibration.

Table 10: Long-list neighborhood strategies with a fixed SIVA model on CUHK-PEDES. Each entry reports R1 / mAP (%). The middle column omits the confidence factor in score calibration, while the last column uses Eq. 5.

Long-list strategy	Association	+ Confidence calibration
SIVA only	79.31 / 70.10	–
DBSCAN (RIVA default)	79.45 / 74.36	79.84 / 76.02
kNN neighborhood	79.37 / 73.81	79.71 / 75.69
k-reciprocal neighborhood	79.48 / 73.97	79.82 / 75.85
Graph diffusion	**79.53 / 74.44**	**79.87 / 76.12**

All four strategies improve mAP over SIVA only, from 70.10% to 73.81%–74.44%. Confidence-guided calibration further improves mAP to 75.69%–76.12%. The final results are close across strategies. Thus, long-list association is not restricted to DBSCAN: it combines candidate neighborhoods with SIVA verification and confidence-guided calibration. DBSCAN provides a simple default with adaptive-size neighborhoods and singleton clusters.

For kNN, we use $k = 3$. The k-reciprocal variant uses the anchor's connected component in a distance-gated reciprocal graph, with $k_r = 3$ and edge-distance threshold $\pi_r = 0.35$. Graph diffusion uses kNN-graph size $k_g = 4$, edge-distance threshold $\pi_e = 0.33$, Gaussian bandwidth $\sigma = 0.33$, diffusion coefficient $\alpha = 0.85$, and output-score threshold $\pi_f = 0.4$. It retains nodes whose normalized diffusion scores exceed π_f. The reciprocal-graph and diffusion variants return a singleton when no other candidate is retained. DBSCAN uses radius 0.33 and $\mathrm{MinPts} = 1$, as in the main experiments. Query expansion is evaluated separately in Table 9, since it updates the query representation rather than constructing an anchor-conditioned visual neighborhood.

E.5 Generalization across Multi-Image MLLMs

We evaluate four native multi-image MLLMs as SIVA backbones using the same RSTPReid listwise samples, three-stage dual-task curriculum, instruction format, and evaluation protocol. Table 11 compares short-list reranking before and after SIVA training; these results do not include the long-list module.

Table 11: SIVA backbone comparison on RSTPReid. Entries report R1 / mAP (%); gains are percentage-point differences after training.

MLLM backbone	Before training	After training	Gain
LLaVA-OneVision-7B	66.85 / 53.87	72.85 / 56.28	+6.00 / +2.41
Qwen2-VL-7B	66.95 / 54.14	73.00 / 56.46	+6.05 / +2.32
Qwen2.5-VL-7B	66.70 / 53.92	73.05 / 56.42	+6.35 / +2.50
Qwen3-VL-8B	67.35 / 54.39	**73.30 / 56.57**	+5.95 / +2.18

Training improves Rank-1 by 5.95–6.35 points and mAP by 2.18–2.50 points across these backbones. These consistent gains demonstrate the effectiveness of SIVA across native multi-image MLLMs.

E.6 Effect of Synthetic Pretraining

We examine the contribution of HAM-PEDES by removing Stage-1 while retaining the target-dataset adaptation and hard-sample refinement stages on RSTPReid. Table 12 reports short-list SIVA performance.

Table 12: Effect of Stage-1 synthetic pretraining on RSTPReid (%).

Training configuration	R1	mAP
Off-the-shelf Qwen2.5-VL-7B	66.70	53.92
Stage-2 only	71.25	55.50
Stage-2 + Stage-3	72.50	56.08
Stage-1 + Stage-2 + Stage-3	**73.05**	**56.42**

Without HAM-PEDES, Stages-2 and -3 improve the untuned MLLM by 5.80 Rank-1 and 2.16 mAP points. Stage-1 provides a further gain of 0.55 and 0.34 points, respectively. Synthetic pretraining is therefore useful but not indispensable, providing an initialization before adaptation to human descriptions and hard examples.

E.7 Evaluation on UFine6926

UFine6926, introduced as part of UFineBench [69], provides more detailed pedestrian descriptions than the three standard benchmarks. We additionally evaluate RIVA on this dataset with ICL as the base retriever. Table 13 shows that RIVA improves ICL by 1.66 Rank-1, 3.27 mAP, and 5.26 mINP points. It also exceeds THI by 0.93, 2.42, and 4.51 points, respectively.

Table 13: Retrieval performance on UFine6926 (%).

Method	R1	mAP	mINP
ICL	91.03	89.58	84.92
ICL + THI	91.76	90.43	85.67
ICL + RIVA	**92.69**	**92.85**	**90.18**

F Repeatability and Deployment Scope

F.1 Repeated Training Runs

We perform three end-to-end training runs for HAM + RIVA on CUHK-PEDES under the same training configuration. Each run starts from the off-the-shelf Qwen2.5-VL checkpoint, executes all three SIVA training stages, and evaluates the complete RIVA pipeline. Table 14 reports the mean and standard deviation across these runs. The mean improvements over HAM are 2.18 Rank-1, 6.69 mAP, and 11.41 mINP points. The small standard deviations indicate stable improvements across the three runs.

Table 14: Mean $\pm$ standard deviation over three training runs on CUHK-PEDES (%). HAM is fixed across runs.

Method	R1	mAP	mINP
HAM	77.63	69.24	53.36
HAM + RIVA	79.81 ± 0.13	75.93 ± 0.37	64.77 ± 0.82

F.2 Inference Repeatability

We repeat RIVA inference on ICFG-PEDES five times with the same checkpoint, inputs, and random seed, obtaining identical retrieval results. Across five runs with different seeds, the parsed image selections agree with the reference run in 99.4% of cases, while the ranges of Rank-1, mAP, and mINP are 0.80, 0.92, and 1.07 percentage points, respectively.

F.3 Inference Cost and Scope

Under the inference setting in Section 4.5, the peak allocated memory per GPU is 23.7 GiB for THI and 23.1 GiB for RIVA. As RIVA only reranks the retrieved top-K candidates, its performance remains limited by the initial retrieval results. Ambiguous descriptions and incorrect anchor selection may also affect the reranking performance.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe candidate-side visual association for TIReID reranking. The main experiments and Appendix E evaluate its effectiveness across different retrieval models, datasets, and MLLM backbones.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion and Appendix F.3 discuss inference cost and the limitations imposed by initial candidate retrieval. Further experiments on different retrieval models and MLLM backbones are provided in Section 4.4 and Appendix E.5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This is an empirical study. The equations define the retrieval and reranking procedures.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [No]

   Justification: We describe the method, training stages, principal hyperparameters, and evaluation protocol in the paper and appendices. Code, checkpoints, and reproduction instructions will be released through the project repository.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We use existing TIReID datasets and describe the training-data construction procedure. Code, checkpoints, and reproduction instructions will be released through the project repository.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [No]

   Justification: The main paper and appendices specify the data splits, training stages, learning rates, list lengths, and reranking parameters. The project repository will provide the training and evaluation configurations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section 4.5 reports the mean and standard deviation across three timing runs. Appendix F reports the mean and standard deviation across three end-to-end training runs for HAM + RIVA on CUHK-PEDES.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: SIVA training uses eight RTX 3090 GPUs. The inference comparison uses two RTX 3090 GPUs and reports wall time, query throughput, and per-GPU peak allocated memory.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work uses existing research datasets and does not introduce new data collection or direct interaction with human participants.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: We discuss beneficial applications and encourage responsible use in the conclusion, but do not provide an analysis of privacy risks or potential misuse.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: The study uses existing research datasets for person retrieval. The conclusion encourages responsible use of the proposed framework.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: We cite the original sources of the datasets and pretrained models used in this study.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: We describe the proposed training-data construction and SIVA model. Code, checkpoints, and accompanying documentation will be released through the project repository.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve new crowdsourcing or research with human participants; it uses existing datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve new data collection or direct interaction with human participants. It uses existing research datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Section 3 describes the MLLM-based SIVA component, its two instruction tasks, and its three-stage training procedure. Appendix E.5 reports the evaluated multi-image MLLM backbones.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.