Chat-Driven Text Generation and Interaction for Person Retrieval

Anonymous ACL submission

Abstract

Text-based person search (TBPS) enables the retrieval of person images from large-scale databases using natural language descriptions, offering critical value in surveillance applications. However, a major challenge lies in the labor-intensive process of obtaining high-quality textual annotations, which limits scalability and practical deployment. To address this, we introduce two complementary modules: **Multi-Turn Text Generation (MTG)** and **Multi-Turn Text Interaction (MTI)**. MTG generates rich pseudo-labels through simulated dialogues with multimodal large language models (MLLMs), producing fine-grained and diverse visual descriptions without manual supervision. MTI refines user queries at inference time through dynamic, dialogue-based reasoning, enabling the system to interpret and resolve vague, incomplete, or ambiguous descriptions—characteristics often seen in real-world search scenarios. Together, MTG and MTI form a unified and annotation-free framework that significantly improves retrieval accuracy, robustness, and usability. Extensive evaluations demonstrate that our method achieves competitive or superior results while eliminating the need for manual captions, paving the way for scalable and practical deployment of TBPS systems.

1 Introduction

Text-based person search (TBPS) aims to retrieve images of a target individual from large-scale galleries using natural language descriptions (Li et al., 2017a). It lies at the intersection of image-text retrieval (Lei et al., 2022; Sun et al., 2021; Miech et al., 2021) and image-based person re-identification (Re-ID) (He et al., 2021; Luo et al., 2019; Wang et al., 2022a), offering a flexible alternative to visual queries. Text queries are more accessible and often provide richer semantic cues about identity, enabling applications ranging from personal photo organization to public security and surveillance.

Since the seminal introduction of CUHK-PEDES (Li et al., 2017a), TBPS has made substantial progress, largely driven by advances in cross-modal representation learning that align visual and textual modalities in a shared embedding space (Radford et al., 2021). However, despite these technical developments, one fundamental bottleneck remains: the reliance on high-quality textual annotations. While visual data can be easily acquired from surveillance footage, generating accurate and semantically rich descriptions is labor-intensive, expensive, and inherently unscalable.

Automated captioning methods provide a partial solution, but often suffer from semantic drift, repetitive phrasing, and hallucinated content (Kolouju et al., 2025), leading to vague or misleading labels (see Figure 1). This limitation motivates a central research question: *Can TBPS be achieved effectively without depending on manually crafted descriptions?*

To address this challenge, we propose **CTGI** (Chat-Driven Text Generation and Interaction), a unified and annotation-free framework designed to bridge the supervision gap through multimodal dialogue. CTGI comprises two synergistic modules: **Multi-Turn Text Generation (MTG)** for training supervision and **Multi-Turn Text Interaction (MTI)** for inference-time query refinement (see Figure 2).

The **MTG** module simulates multi-turn conversations with an MLLM to generate rich pseudo-labels. Starting from a baseline caption, it iteratively refines the description using a series of attribute-targeted prompts that mimic human dialogue. This process leads to semantically dense, diverse, and fine-grained annotations that far exceed the quality of single-turn captioning. To accommodate these longer descriptions, we extend CLIP's default 77-token input limit by applying po-

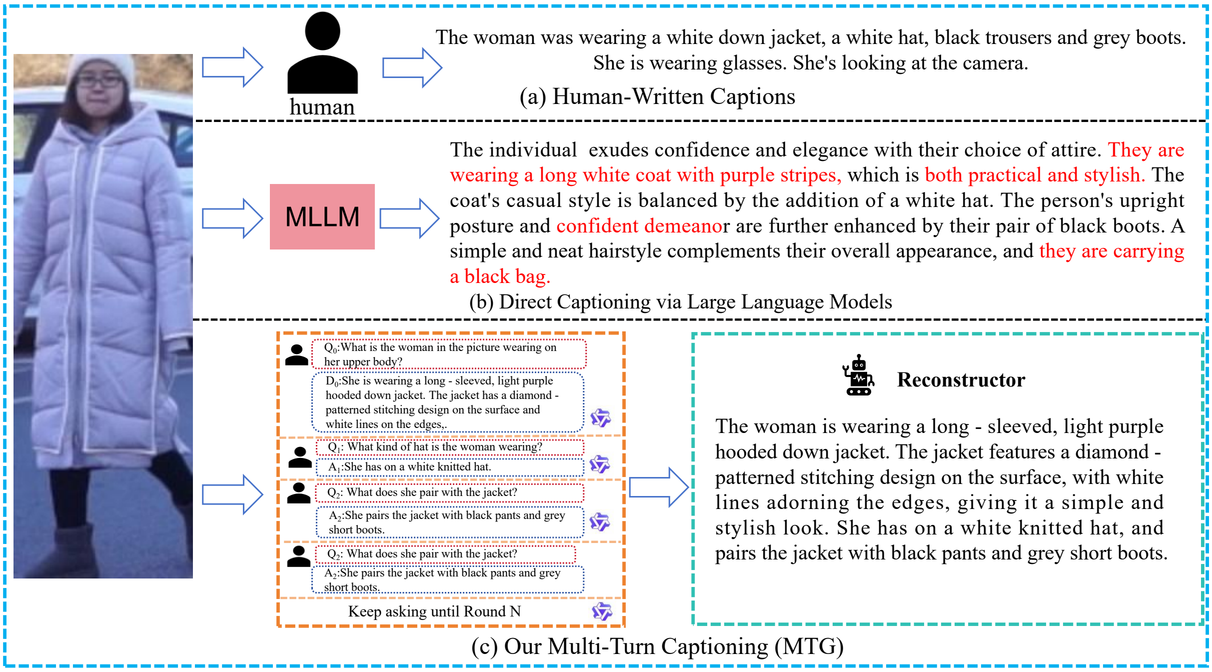

Figure 1: Comparison of person description strategies. (a) Human-written captions are concise but often lack compositional depth and attribute coverage. (b) Direct captioning with large language models (LLMs) generates descriptions in a single forward pass, but often suffers from hallucinations or omissions—particularly in capturing fine-grained visual details such as clothing, accessories, or scene context. (c) Our proposed multi-turn strategy simulates an interactive dialogue with the MLLM, progressively enriching descriptions through targeted Q&A, yielding more expressive, accurate, and human-aligned captions.

sitional embedding stretching—retaining the first 20 learned positions and interpolating the remaining embeddings to support up to 248 tokens without retraining the model.

The **MTI** module operates during inference to refine under-specified user queries through MLLM-driven dialogue. It begins by identifying a candidate anchor image and then generates targeted questions to extract missing or ambiguous attributes. The responses are aggregated into a refined query that is better aligned with the target image. MTI also incorporates filtering mechanisms to avoid redundancy and maintain efficiency. As a plug-and-play module, MTI can be easily deployed with various pretrained vision-language retrieval models with minimal adaptation cost.

Our key contributions are as follows:

- We propose **CTGI**, a novel chat-driven framework for TBPS that eliminates the need for manual annotations by unifying pseudo-caption generation and interactive query refinement.

- We develop **MTG**, a multi-turn captioning module that generates rich, attribute-aware pseudo-labels through iterative dialogue, and supports long-text encoding via positional embedding extension.

- We introduce **MTI**, a dynamic inference module that refines natural language queries via MLLM-guided interaction, enhancing alignment between user input and visual content for more accurate retrieval.

2 Related Work

Text-Based Person Search (TBPS) has progressed significantly since the release of CUHK-PEDES (Li et al., 2017a). Early efforts focused on embedding visual and textual data into a shared space, evolving from global alignment (Zheng et al., 2020; Farooq et al., 2020) to fine-grained matching (Chen et al., 2018, 2022; Suo et al., 2022), often enhanced by pose cues (Jing et al., 2020), part-level features (Wang et al., 2020), or semantic knowledge (Loper and Bird, 2002). In parallel, representation learning approaches aimed to extract modality-invariant features by addressing background clutter (Zhu et al., 2021a), color sensitivity (Wu et al., 2021), and multi-scale fusion (Shao et al., 2022). Recently, large-scale pretrained models like CLIP (Radford et al., 2021) have enabled

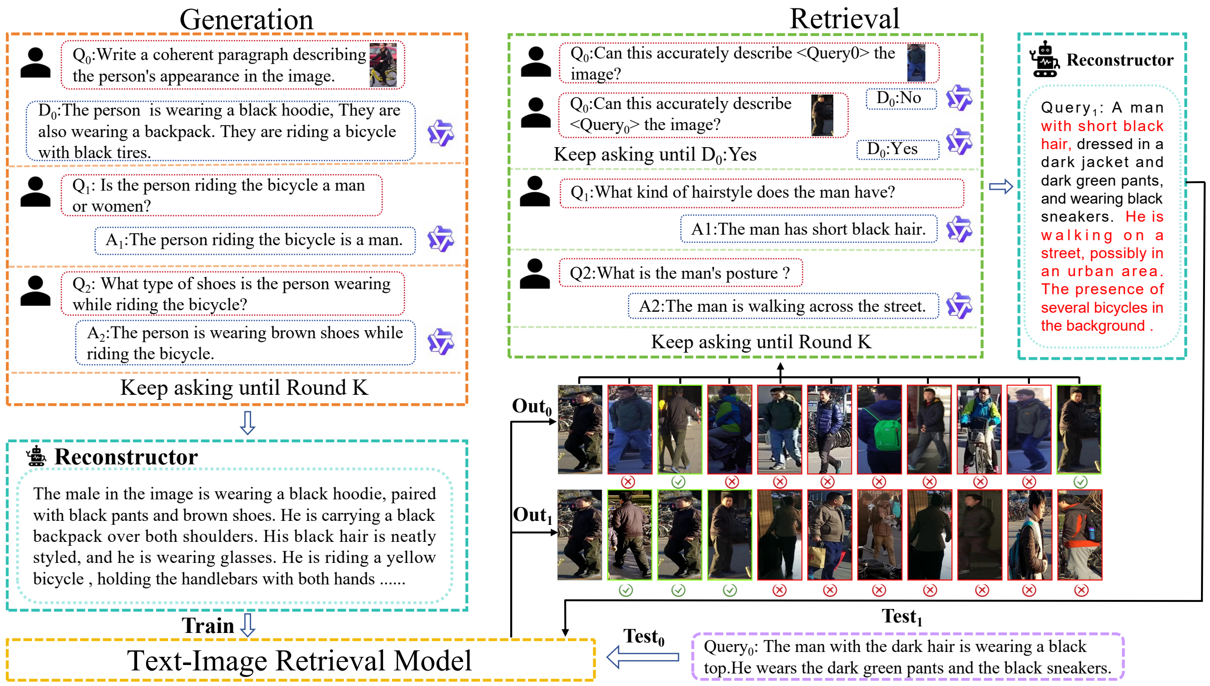

Figure 2: Overview of the proposed **CTGI** framework for text-based person search. The framework consists of two stages: (1) **Training-time generation**: MTG simulates multi-turn dialogue to iteratively enrich captions, while a reconstructor synthesizes pseudo-labels using structured prompts; and (2) **Inference-time retrieval**: MTI refines user queries through MLLM-driven Q&A, enhancing alignment between the query and candidate images for improved re-ranking.

strong generalization for cross-modal retrieval with minimal tuning (Jiang and Ye, 2023b; Han et al., 2021; Wei et al., 2023), with IRRA (Jiang and Ye, 2023b) improving alignment via multimodal interaction.

Despite these advances, most TBPS methods still depend on costly human-annotated text, limiting scalability. Weakly supervised (Zhao et al., 2021) and synthetic labeling (Yang et al., 2023; Tan et al., 2024) offer partial relief but struggle with vague or conversational queries.

To overcome this, we propose a new TBPS paradigm—**Chat-Driven Text Generation and Interaction (CTGI)**—which eliminates manual annotations and enhances retrieval through multi-turn dialogue with Multimodal Large Language Models (MLLMs). Unlike earlier interactive retrieval systems (Guo et al., 2018; Lee et al., 2024) that require task-specific data or retraining, CTGI supports open-ended, behavior-centric queries and dynamically refines both pseudo-labels and user input. By leveraging MLLMs as plug-and-play agents, CTGI achieves robust, scalable, and annotation-free TBPS—bridging the gap between lab settings and real-world deployments.

3 Methodology

In this section, we briefly outline a **C**hat-Driven **T**ext **G**eneration and **I**nteraction (**CTGI**) model for person retrieval. The CTGI model framework consists of two main modules: (1) The Multi-Turn Text Generation (**MTG**) module, which uses a multimodal large language model to generate detailed textual descriptions for given person images through an interactive Q&A dialogue; and (2) The Multi-Turn Text Interaction (**MTI**) module, which is used in an inference-time pipeline that refines the textual query by leveraging visual context from retrieved images and then performs re-ranking. The overall framework is illustrated in Figure 2.

3.1 Multi-Turn Text Generation

The Multi-Turn Text Generation module generates a comprehensive pseudo-label for each person image I by iteratively querying a multimodal large language model for fine-grained details. This process is initiated with an initial captioning prompt designed to elicit a general description. Given an image I, we can use the MLLM with a prompt P_{init}, *e.g.*, *"Describe the person in the image,"* yielding an initial static caption T_s:

$$T_s = MLLM\big(I, P_{\text{init}}\big), \tag{1}$$

However, T_s provides only a simple, basic textual description and often overlooks distinctive attributes. To capture more distinctive attributes of a person, the QA-guided refinement rounds method provides a more detailed textual description improvement strategy. Specifically, in each round i, the model generates an answer a_i that aligns with the image content based on a specific question q_i, *e.g.*,

q_i: *Is the person riding the bicycle?*
a_i: *Yes, the person is riding the bicycle.*

After N rounds of QA operations, we obtain all preceding QA results $\{(q_i, a_i)\}_{i=1}^{N}$ concatenated together to obtain the enriched caption T_e:

$$T_e = MLLM\big([a_1, a_2, ..., a_N]\big), \tag{2}$$

Compared to T_s, T_e provides more fine-grained attributes for the given person image, *e.g.*, colors, clothing details, and physical features, which greatly enhance the textual description.

It is important to note that due to the presence of similar questions in the question list, this may lead to repetitive answers. To remove the redundant descriptions, we use the MLLM again and reconstruct T_e by incorporating T_s:

$$T_e = MLLM\big(T_e \mid T_s, p\big), \tag{3}$$

where p denotes the input prompt to the MLLM, *e.g., "Rephrase the description using all the above information."* Compared to T_e in Eq. (2), T_e in Eq. (3) provides a more concise and effective textual description, rather than increasing the quantity of image-related details. Meanwhile, compared to T_s in Eq. (1), T_e contains more details extracted during the MLLM Q&A process, and better aligns with human attention to core image information.

3.2 More Text Positional Embeddings

CLIP's original 77-token limit, imposed by its fixed-length absolute positional embeddings, restricts its ability to process long and detailed text—a critical limitation for tasks such as Text-Based Person Search (TBPS). To address this, we adopt a knowledge-preserving *positional embedding stretching* technique that extends the model's input capacity while maintaining compatibility with pretrained weights.

Following Long-CLIP (Zhang et al., 2024)and FineLIP(Asokan et al., 2025), we preserve the first 20 learned positional embeddings, which are empirically the most well-trained, and interpolate the remaining positions (21–77) to reach a new input length of 248 tokens by applying a 4× stretching factor.

Let $PE(pos)$ denote the original positional embedding at position $pos \in [1, 77]$. We construct the stretched embedding $PE^*(pos)$ for the extended range $pos \in [1, 248]$ as:

$$PE^*(pos) = \begin{cases} PE(pos), & \text{for } pos \le 20 \\ (1-\alpha) \cdot PE\left(\left\lfloor \frac{pos}{\lambda_2} \right\rfloor\right) & \\ \quad +\alpha \cdot PE\left(\left\lceil \frac{pos}{\lambda_2} \right\rceil\right), & \text{for } 21 \le pos \le 77 \end{cases} \tag{4}$$

Here, $\lambda = \frac{248-20}{77-20} \approx 4$ is the interpolation factor, and α is the fractional part of $\frac{pos-20}{\lambda}$. This ensures smooth interpolation while preserving pretrained embeddings for the initial positions.

Inspired by LiT (Zhai et al., 2022), this approach avoids reinitialization or retraining, and allows CLIP to encode longer, semantically rich descriptions generated by the MTG module. Empirical results in Table 4 confirm that this strategy enhances retrieval performance without sacrificing alignment learned during pretraining.

3.3 Multi-Turn Text Interaction (MTI)

MTI operates during inference to resolve underspecified or vague user queries through multi-turn interaction.

Step 1: Anchor Identification. Given a user query q, the system retrieves top-K candidates $\{\hat{v}_1, ..., \hat{v}_K\}$ using similarity score $S_{q,v}$. For each $\hat{v}_k$, the MLLM is prompted to judge alignment with q. The first affirmative response identifies the anchor $\bar{v}$. If no match is found within K attempts, no refinement is applied.

Step 2: Interactive Refinement. With anchor $\bar{v}$, MTI generates a diagnostic question set $\{c_i\}$ focused on missing attributes. Responses are obtained via visual Q&A:

$$r_{\bar{v}} = \text{MLLM}(T_{\text{vqa}}(\{c_i\}, \bar{v})) \tag{5}$$

The final query $\hat{q}$ is synthesized using a template prompt to merge $r_{\bar{v}}$ and q:

$$\hat{q} = \text{MLLM}(T_{\text{aggr}}(r_{\bar{v}}, q)) \tag{6}$$

Step 3: Re-ranking. The final similarity is computed as:

$$\hat{S}_{q,v} = \lambda S_{q,v} + (1 - \lambda) S_{\hat{q},v} \quad (7)$$

with $\hat{S}_{q,\bar{v}} = 1$ to promote anchor matching. Early stopping is triggered when $\hat{v}_1$ surpasses threshold $\xi = 0.85$.

3.4 Reconstructor

The **Reconstructor** plays a pivotal role in transforming fragmented outputs from multi-turn Q&A into coherent and high-quality descriptions. It is deployed in both training and inference pipelines to enhance the effectiveness of CTGI without requiring any manual annotations or dataset-specific tuning.

To ensure the quality of generated descriptions during training, MTG maintains a dynamic question pool and discards Q&A pairs that exhibit low semantic relevance or redundant information. This filtering helps avoid overlong or repetitive captions.

For synthesis, the Reconstructor leverages the **GPT-4o API** to convert structured Q&A logs into fluent and semantically rich pseudo-captions. These refined captions serve as supervision signals for training downstream retrieval models.

In the inference stage, the Reconstructor also contributes to query refinement within MTI. A set of curated diagnostic templates (e.g., "Is the person wearing a backpack?") is used to identify typical ambiguities. These templates help elicit missing attributes without introducing generic or noisy questions. The responses are then aggregated into a revised query that is semantically aligned with the visual anchor.

This unified design ensures that CTGI can support both training-time pseudo-label generation and test-time query refinement effectively—without reliance on human-written descriptions or task-specific engineering.

4 Experiments

We evaluate our framework by re-annotating three public datasets with enriched textual descriptions that offer greater semantic depth and diversity. We compare retrieval models trained on these pseudo-labels against those trained on original annotations. To test generalizability, we integrate our method into standard TBPS pipelines and assess its impact. Finally, we perform ablation studies and visual analyses to better understand the method's effectiveness.

4.1 Datasets and Performance Measurements

We evaluate our approach using three Text-based Person Retrieval datasets: CUHK-PEDES (Li et al., 2017b), ICFG-PEDES (Ding et al., 2021b), and RSTPReid (Zhu et al., 2021b). Our training solely utilizes image data, devoid of any dependency on manually annotated text data. During the testing phase, captions from the dataset are leveraged for re- trieval.

Evaluation Metrics. Following standard practice, we evaluate using Rank-k (k=1,5,10), mean Average Precision (mAP). Higher values indicate better retrieval performance.

4.2 Implementation Details

We evaluate **CTGI** using two strong TBPS baselines: **IRRA** (Jiang and Ye, 2023a) and **RDE** (Qin et al., 2024), both built on **CLIP-ViT/B-16** (Radford et al., 2021). For multimodal reasoning, we adopt **Qwen2-VL-7B** (Wang et al., 2024) as the core MLLM, while the **Reconstructor** leverages the **OpenAI GPT-4o API** (OpenAI, 2023) for pseudo-caption synthesis.

All models follow the original training setups of IRRA and RDE. Input images are resized to 384×128, and standard augmentations (flip, crop, erase) are applied. To support longer text, we extend CLIP's 77-token limit to **248 tokens** by preserving the first 20 positional embeddings and interpolating the rest $4\times$, following (Zhai et al., 2022). The learning rate is set to 1×10^{-5} (with 5 warmup epochs from 1×10^{-6}), and 5×10^{-5} for randomly initialized layers. Cosine decay is used throughout 60 training epochs.

During training, the **MTG** module runs **6 Q&A rounds** per image to generate dense pseudo-labels. For inference, **MTI** examines the top $K = 20$ retrieval candidates, and early exits if the top-1 similarity exceeds $\xi = 0.85$ and is confirmed by the MLLM. Final retrieval scores are fused via weighted re-ranking. All experiments are conducted on **2**$\times$ **NVIDIA RTX 4090 GPUs** with generation temperature fixed at **0.01** for stability.

4.3 Comparison with the State-of-the-Art

We evaluate the effectiveness of our proposed CTGI framework on three widely used benchmark datasets for text-based person search, comparing against both unsupervised and fully supervised

Table 1: Performance on CUHK-PEDES . *: trained with LLaVA-1.5 captions. The best and second-best results are in **bold** and <u>underline</u>, respectively.

Methods	Ref.	Image Enc.	Text Enc.	R-1	R-5	R-10	mAP
Fully Supervised							
TIMAM (Sarafianos et al., 2019)	ICCV'19	RN101	BERT	54.51	77.56	79.27	-
ViTAA (Wang et al., 2020)	ECCV'20	RN50	LSTM	54.92	75.18	82.90	51.60
NAFS (Gao et al., 2021)	arXiv'21	RN50	BERT	59.36	79.13	86.00	54.07
DSSL (Zhu et al., 2021a)	ACMMM'21	RN50	BERT	59.98	80.41	87.56	-
SSAN (Ding et al., 2021a)	arXiv'21	RN50	LSTM	61.37	80.15	86.73	-
Lapscore (Wu et al., 2021)	ICCV'21	RN50	BERT	63.40	-	87.80	-
ISANet (Yan et al., 2022b)	arXiv'22	RN50	LSTM	63.92	82.15	87.69	-
SAF (Li et al., 2022)	ICASSP'22	ViT-Base	BERT	64.13	82.62	88.40	-
DCEL (Qin et al., 2022)	ACMMM'22	CLIP-ViT	CLIP-Xformer	71.36	88.11	92.48	64.25
IVT (Shu et al., 2022)	ECCVW'22	ViT-Base	BERT	65.59	83.11	89.21	-
CFine (Yan et al., 2022a)	TIP'23	CLIP-ViT	BERT	69.57	85.93	91.15	-
IRRA (Jiang and Ye, 2023c)	CVPR'23	CLIP-ViT	CLIP-Xformer	73.38	89.93	93.71	66.13
BiLMa (Fujii and Tarashima, 2023)	ICCV'23	CLIP-ViT	CLIP-Xformer	74.03	89.59	93.62	66.57
PBSL (Shen et al., 2023)	ACMMM'23	RN50	BERT	65.32	83.81	89.26	-
BEAT (Ma et al., 2023)	ACMMM'23	RN101	BERT	65.61	83.45	89.54	-
LCR^{2}S (Yan et al., 2023)	ACMMM'23	RN50	TextCNN	67.36	84.19	89.62	59.24
DCEL (Li et al., 2023)	ACMMM'23	CLIP-ViT	CLIP-Xformer	75.02	90.89	94.52	-
UniPT (Shao et al., 2023)	ICCV'23	CLIP-ViT	CLIP-Xformer	68.50	84.67	-	-
TBPS (Cao et al., 2024)	AAAI'24	CLIP-ViT	CLIP-Xformer	73.54	88.19	92.35	65.38
RDE (Qin et al., 2024)	CVPR'24	CLIP-ViT	CLIP-Xformer	75.94	90.14	94.12	67.56
CFAM (Zuo et al., 2024)	CVPR'24	CLIP-ViT	CLIP-Xformer	75.60	90.53	-	67.27
MLLM+IRRA (Wentao Tan, 2024)	CVPR'24	CLIP-ViT	CLIP-Xformer	76.82	91.16	-	69.55
MGRL (Lv et al., 2024)	ICASSP'24	CLIP-ViT	CLIP-Xformer	73.91	90.68	-	67.28
OCDL (Li et al., 2025a)	ICASSP'25	CLIP-ViT	CLIP-Xformer	75.10	89.43	-	68.18
Unsupervised							
IRRA* (Li et al., 2025b)	CVPR'23	CLIP-ViT	CLIP-Xformer	32.94	54.37	64.67	30.87
BLIP* (Li et al., 2025b)	ICML'22	BLIP-ViT	BLIP-Xformer	51.41	71.41	78.76	44.73
GTR (Bai et al., 2023)	MM'23	BLIP-ViT	BLIP-Xformer	47.53	68.23	75.91	42.91
MUMA (Li et al., 2025b)	AAAI'25	BLIP-ViT	BLIP-Xformer	59.52	77.79	-	52.75
Our+IRRA	-	CLIP-ViT	CLIP-Xformer	<u>63.53</u>	<u>80.25</u>	<u>87.84</u>	<u>52.37</u>
Our+RDE	-	CLIP-ViT	CLIP-Xformer	**67.82**	**85.45**	**90.63**	**55.14**

state-of-the-art methods. Our framework is instantiated with two variants, *Our+IRRA* and *Our+RDE*, which employ different retrieval backbones while sharing the same underlying CTGI components.

CUHK-PEDES: As reported in Table 1, under the unsupervised setting, our *Our+RDE* achieves a Rank-1 of 67.82% and mAP of 55.14%, substantially outperforming the strongest unsupervised baseline MUMA, which obtains 59.52% and 52.75% respectively. Notably, *Our+IRRA* also surpasses MUMA by a clear margin, demonstrating the strong efficacy of CTGI in generating informative pseudo-labels and improving retrieval without manual annotations. Compared with fully supervised methods, our results approach competitive levels, surpassing several mid-tier supervised models and narrowing the gap to the top performers.

ICFG-PEDES: Table 2 shows that our framework maintains state-of-the-art performance in the unsupervised category with a Rank-1 of 56.16% and mAP of 32.40% for *Our+RDE*, exceeding the best supervised methods in some metrics. This highlights CTGI's robustness and generalization ability across datasets with different granularity and annotation styles. The improvements over other unsupervised baselines such as BLIP and GTR further confirm the superiority of our approach.

RSTPReid: As shown in Table 3, on the RSTPReid dataset, *Our+RDE* achieves a Rank-1 of 66.35% and mAP of 51.51%, outperforming the second-best unsupervised method MUMA by approximately 12% in Rank-1 and over 11% in mAP.

Table 2: Performance on ICFG-PEDES. *: trained with LLaVA-1.5 captions.The best and second-best results are in **bold** and <u>underline</u>, respectively.

Method	R@1	R@5	R@10	mAP
Fully Supervised				
Dual Path (Zheng et al., 2020)	38.99	59.44	68.41	-
CMPM/C (Zhang and Lu, 2018)	43.51	65.44	74.26	-
ViTAA (Wang et al., 2020)	50.98	68.79	75.78	-
SSAN (Ding et al., 2021a)	54.23	72.63	79.53	-
IVT (Shu et al., 2022)	56.04	73.60	80.22	-
ISANet (Yan et al., 2022b)	57.73	75.42	81.72	-
CFine (Yan et al., 2022a)	60.83	76.55	82.42	-
IRRA (Jiang and Ye, 2023c)	63.46	80.25	85.82	38.06
BiLMa (Fujii and Tarashima, 2023)	63.83	80.15	85.74	38.26
PBSL (Shen et al., 2023)	57.84	75.46	82.15	-
BEAT (Ma et al., 2023)	58.25	75.92	81.96	-
LCR^{2}S (Yan et al., 2023)	57.93	76.08	82.40	38.21
DCEL (Li et al., 2023)	64.88	81.34	86.72	-
UniPT (Shao et al., 2023)	60.09	76.19	-	-
TBPS (Cao et al., 2024)	65.05	80.34	85.47	39.83
CFAM (Zuo et al., 2024)	65.38	81.17	-	39.42
MGRL (Lv et al., 2024)	67.28	63.87	-	82.34
OCDL (Li et al., 2025a)	64.53	80.23	-	40.76
Unsupervised				
IRRA* (Li et al., 2025b)	21.23	37.37	46.04	11.47
BLIP* (Li et al., 2025b)	31.58	52.03	61.73	13.20
GTR (Bai et al., 2023)	28.25	45.21	53.51	13.82
MUMA (Li et al., 2025b)	38.11	56.01	63.96	19.02
Ours + IRRA	<u>48.76</u>	<u>67.38</u>	<u>74.66</u>	<u>27.42</u>
Ours + RDE	**56.16**	**73.18**	**79.42**	**32.40**

Table 3: Performance on RSTPReid. *: trained with LLaVA-1.5 captions.The best and second-best results are in **bold** and <u>underline</u>, respectively.

Methods	R-1	R-5	R-10	mAP	
Fully Supervised					
DSSL (Zhu et al., 2021a)	39.05	62.60	73.95	-	
SSAN (Ding et al., 2021a)	43.50	67.80	77.15	-	
LBUL (Wang et al., 2022b)	45.55	68.20	77.85	-	
IVT (Shu et al., 2022)	46.70	70.00	78.80	-	
CFine (Yan et al., 2022a)	50.55	72.50	81.60	-	
IRRA (Jiang and Ye, 2023c)	60.20	81.30	88.20	47.17	
BiLMA (Fujii and Tarashima, 2023)	61.20	81.50	88.80	48.51	
PBSL (Shen et al., 2023)	47.80	71.40	79.90	-	
BEAT (Ma et al., 2023)	48.10	73.10	81.30	-	
LCR^{2}S (Yan et al., 2023)	54.95	76.65	84.70	40.92	
DCEL (Li et al., 2023)	61.35	83.95	90.45	-	
TBPS (Cao et al., 2024)	61.95	83.55	88.75	48.26	
CFAM (Zuo et al., 2024)	62.45	83.55	-	49.50	
OCDL (Li et al., 2025a)	61.60	82.35	-	49.77	
Unsupervised					
IRRA* (Li et al., 2025b)	37.60	60.65	72.30	27.42	-
BLIP* (Li et al., 2025b)	44.45	67.70	77.25	33.73	-
GTR (Bai et al., 2023)	45.60	70.35	79.95	33.30	
MUMA (Li et al., 2025b)	54.35	76.05	83.65	40.50	
Our+IRRA	<u>64.20</u>	<u>83.55</u>	<u>90.30</u>	<u>49.66</u>	
Our+RDE	**66.35**	**85.50**	**91.24**	**51.51**	

Moreover, our method exceeds the performance of several fully supervised models, including CFine, illustrating the strong competitiveness and scalability of CTGI without reliance on any manual annotations.

Across all datasets, our CTGI framework demonstrates a consistent and significant improvement over existing unsupervised methods, closing the gap towards fully supervised performance. These results validate the effectiveness of leveraging multimodal large language models for pseudo-label generation and interactive query refinement, enabling robust and scalable text-based person search in practical scenarios.

4.4 Ablation Study

We conduct ablation experiments on the RSTPReid dataset to systematically analyze the individual and combined effects of Multi-Turn Text Generation (MTG) and Multi-Turn Text Interaction (MTI). When employed separately, MTG enhances retrieval by generating detailed and semantically rich pseudo-labels, resulting in notable improvements in Rank-1 accuracy and mAP over the baseline. For instance, with the IRRA backbone, MTG alone achieves a Rank-1 of 52.30%, indicating its strong ability to provide effective training supervision through enriched textual descriptions.

Similarly, MTI, which refines user queries at inference time via multi-turn dialogue, independently boosts performance by improving the semantic alignment between queries and visual features. This is reflected by an increased Rank-1 accuracy of 55.50% with IRRA, highlighting MTI's effectiveness in mitigating ambiguity in free-form textual queries.

Importantly, the integration of MTG and MTI yields complementary benefits, producing the highest gains across all metrics. Combined, they achieve Rank-1 accuracies of 64.20% and 66.35% with IRRA and RDE backbones respectively, alongside corresponding mAP improvements. These results confirm that the synergy between richer pseudo-label generation and dynamic query refinement substantially advances cross-modal retrieval performance and robustness.

Table 4: Ablation study on the RSTPReid dataset. MTG: Multi-Turn Text Generation, MTI: Multi-Turn Text Interaction, PES: Positional Embedding Stretching.

Method	MTG	MTI	Rank-1	Rank-5	Rank-10	mAP
Our+IRRA	✓		52.30	74.65	84.05	40.03
Our+IRRA		✓	55.50	77.50	86.55	44.87
Our+IRRA	✓	✓	64.20	83.55	90.30	48.03
Our+IRRA (w/o PES)	✓	✓	63.00	82.65	88.80	47.60
Our+RDE	✓		60.55	79.85	86.30	44.98
Our+RDE		✓	62.55	82.85	89.00	46.43
Our+RDE	✓	✓	66.35	85.50	91.25	49.66
Our+RDE (w/o PES)	✓	✓	65.75	84.05	90.60	49.60

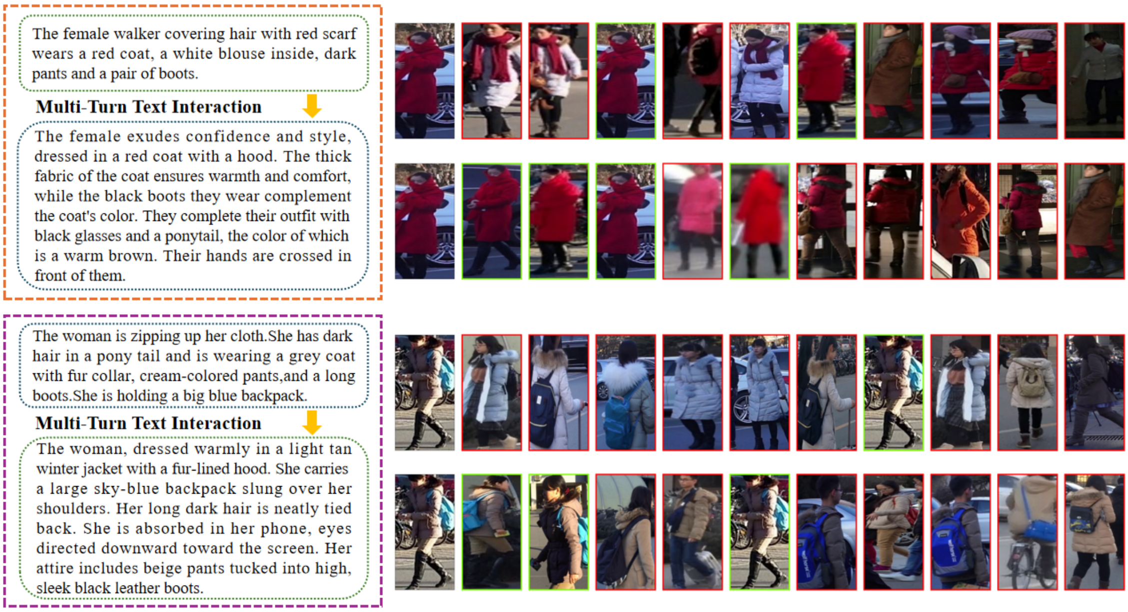

Figure 3: Top-10 retrieval results on the RSTPReid dataset. The first column is the ground-truth image. The first row shows retrieval results using IRRA; the second row shows results after applying IRRA with MTI. Refined queries generated by multi-turn interaction are shown alongside each example. Green borders indicate correct matches.

4.5 Visualization of Retrieval Results

To evaluate the effectiveness of MTI, we conducted controlled experiments with a fixed operation cycle. Figure 3 visualizes the top-10 retrieval results before and after applying MTI. Notably, the retrieval model is trained solely on pseudo-captions generated by the MTG module, without any manual annotations. Due to the incomplete alignment between initial queries and ground-truth test captions, retrieval without MTI often yields suboptimal results. In contrast, MTI dynamically refines the query through interactive optimization, enabling more accurate and robust ranking performance.

5 Conclusion

In this work, we introduced **CTGI** (Chat-Driven Text Generation and Interaction), a unified and annotation-free framework for Text-Based Person Search (TBPS) that removes the dependency on manually crafted textual descriptions. CTGI integrates two synergistic modules: **Multi-Turn Text Generation (MTG)** for training supervision and **Multi-Turn Text Interaction (MTI)** for inference-time refinement. Together, they leverage the expressive capabilities of Multimodal Large Language Models (MLLMs) to generate rich pseudo-labels and iteratively enhance user queries via natural language dialogue. Extensive experiments across multiple TBPS benchmarks show that CTGI achieves competitive or superior performance compared to fully supervised methods, while seamlessly adapting to existing retrieval pipelines. Ablation studies and qualitative visualizations further underscore the value of multi-turn interaction and MLLM-guided refinement in improving cross-modal alignment and retrieval robustness.

Limitations

While **CTGI** demonstrates strong performance without manual annotations, several challenges remain. First, pseudo-labels generated by MTG may contain semantic noise or redundancy. Although robust retrieval backbones like RDE are designed for noisy environments and thus benefit more from such supervision, other models without inherent noise-filtering may be more vulnerable to degraded performance. Second, MTI introduces additional inference overhead due to multi-turn interactions with MLLMs. Even with early stopping and anchor validation, this can limit deployment in latency-sensitive applications. Third, both MTG and MTI rely on the generalization ability of the underlying MLLM (e.g., Qwen2-VL-7B), which may yield suboptimal results in unfamiliar domains or when handling fine-grained attributes. Future work could address these issues through uncertainty-aware label filtering, more efficient MLLMs, and domain-adaptive interaction strategies.

References

Mothilal Asokan, Kebin Wu, and Fatima Albreiki. 2025. Finelip: Extending clip's reach via fine-grained alignment with longer text inputs. *arXiv preprint arXiv:2504.01916*.

Yang Bai, Jingyao Wang, Min Cao, Chen Chen, Ziqiang Cao, Liqiang Nie, and Min Zhang. 2023. Text-based person search without parallel image-text data. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 757–767.

Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. 2024. An empirical study of clip for text-based person search.

Dapeng Chen, Hongsheng Li, Xihui Liu, Yantao Shen, Jing Shao, Zejian Yuan, and Xiaogang Wang. 2018. Improving deep visual representation for person re-identification by global and local image-language association. In *Proceedings of the European conference on computer vision (ECCV)*, pages 54–70.

Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. 2022. Tipcb: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494:171–181.

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

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

Ammarah Farooq, Muhammad Awais, Fei Yan, Josef Kittler, Ali Akbari, and Syed Safwan Khalid. 2020. A convolutional baseline for person re-identification using vision and language descriptions. *arXiv preprint arXiv:2003.00808*.

Takuro Fujii and Shuhei Tarashima. 2023. Bilma: Bidirectional local-matching for text-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2786–2790.

Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. 2021. Contextual non-local alignment over full-scale representation for text-based person search. *arXiv:2101.03036*.

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018. Dialog-based interactive image retrieval. In *Advances in Neural Information Processing Systems*, pages 678–688.

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

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. 2021. Transreid: Transformer-based object re-identification. In *ICCV*, pages 15013–15022.

D. Jiang and M. Ye. 2023a. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2787–2797.

Ding Jiang and Mang Ye. 2023b. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ding Jiang and Mang Ye. 2023c. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *CVPR*, pages 2787–2797.

Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. 2020. Pose-guided multi-granularity attention network for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11189–11196.

Pranavi Kolouju, Eric Xing, Robert Pless, Nathan Jacobs, and Abby Stylianou. 2025. good4cir: Generating detailed synthetic captions for composed image retrieval. *arXiv preprint arXiv:2503.17871*.

Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. 2024. Interactive text-to-image retrieval with large language models: A plug-and-play approach. *arXiv preprint arXiv:2406.03411*.

Jie Lei, Xinlei Chen, Ning Zhang, Mengjiao Wang, Mohit Bansal, Tamara L Berg, and Licheng Yu. 2022. Loopitr: Combining dual and cross encoder architectures for image-text retrieval. *arXiv:2203.05465*.

Haiwen Li, Delong Liu, Fei Su, and Zhicheng Zhao. 2025a. Object-centric discriminative learning for text-based person retrieval. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Shenshen Li, Xing Xu, Yang Yang, Fumin Shen, Yijun Mo, Yujie Li, and Heng Tao Shen. 2023. Dcel: Deep cross-modal evidential learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6292–6300.

Shiping Li, Min Cao, and Min Zhang. 2022. Learning semantic-aligned feature representation for text-based person search. In *ICASSP*, pages 2724–2728. IEEE.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017a. Person search with natural language description. In *CVPR*, pages 1970–1979.

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

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

Zongyi Li, Li Jianbo, Yuxuan Shi, Jiazhong Chen, Shijuan Huang, Linnan Tu, Fei Shen, and Hefei Ling. 2025b. Exploring the potential of large vision-language models for unsupervised text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5119–5127.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. 2019. Bag of tricks and a strong baseline for deep person re-identification. In *CVPR workshops*, pages 0–0.

Tianle Lv, Shuang Li, Jiaxu Leng, and Xinbo Gao. 2024. Mgrl: Mutual-guidance representation learning for text-to-image person retrieval. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2895–2899. IEEE.

Yiwei Ma, Xiaoshuai Sun, Jiayi Ji, Guannan Jiang, Weilin Zhuang, and Rongrong Ji. 2023. Beat: Bi-directional one-to-many embedding alignment for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4157–4168.

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2021. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *CVPR*, pages 9826–9836.

OpenAI. 2023. Gpt-4o. Large language model.

Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. 2024. Noisy-correspondence learning for text-to-image person re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. 2022. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4948–4956.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR.

Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. 2019. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5814–5824.

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

Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. 2022. Learning granularity-unified representations for text-to-image person re-identification. In *ACM MM*, pages 5566–5574.

Fei Shen, Xiangbo Shu, Xiaoyu Du, and Jinhui Tang. 2023. Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8922–8931.

Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. 2022. See finer, see more: Implicit modality alignment for text-based person retrieval. In *European Conference on Computer Vision*, pages 624–641. Springer.

Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. 2021. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *NAACL-HLT*, pages 982–997.

Wei Suo, Mengyang Sun, Kai Niu, Yiqi Gao, Peng Wang, Yanning Zhang, and Qi Wu. 2022. A simple and robust correlation filtering method for text-based person search. In *ECCV*, pages 726–742. Springer.

Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. 2024. Harnessing the power of mllms for transferable text-to-image person reid. In *CVPR*, pages 17127–17137.

Haochen Wang, Jiayi Shen, Yongtuo Liu, Yan Gao, and Efstratios Gavves. 2022a. Nformer: Robust person re-identification with neighbor transformer. In *CVPR*, pages 7297–7307.

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. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhe Wang, Zhiyuan Fang, Jun Wang, and Yezhou Yang. 2020. Vitaa: Visual-textual attributes alignment in person search by natural language. In *ECCV*, pages 402–420. Springer.

Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. 2022b. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *ACM MM*, pages 1984–1992.

Donglai Wei, Sipeng Zhang, Tong Yang, and Jing Liu. 2023. Calibrating cross-modal feature for text-based person searching. *arXiv preprint arXiv:2304.02278*.

Jiayu Jiang Fei Wang Yibing Zhan Dapeng Tao Wentao Tan, Changxing Ding. 2024. Harnessing the power of mllms for transferable text-to-image person reid. *CVPR*.

Yushuang Wu, Zizheng Yan, Xiaoguang Han, Guanbin Li, Changqing Zou, and Shuguang Cui. 2021. Lapscore: language-guided person search via color reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1624–1633.

Shuanglin Yan, Neng Dong, Jun Liu, Liyan Zhang, and Jinhui Tang. 2023. Learning comprehensive representations with richer self for text-to-image person re-identification. In *Proceedings of the 31st ACM international conference on multimedia*, pages 6202–6211.

Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. 2022a. Clip-driven fine-grained text-image person re-identification. *arXiv:2210.10276*.

Shuanglin Yan, Hao Tang, Liyan Zhang, and Jinhui Tang. 2022b. Image-specific information suppression and implicit local alignment for text-based person search. *arXiv preprint arXiv:2208.14365*.

Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. 2023. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *ACM MM*, pages 4492–4501.

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

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer.

Ying Zhang and Huchuan Lu. 2018. Deep cross-modal projection learning for image-text matching. In *ECCV*, pages 686–701.

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

Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(2):1–23.

Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021a. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *ACM MM*, pages 209–217.

Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021b. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*.

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

A Datasets

CUHK-PEDES (Li et al., 2017c) is the first and most widely used dataset for text-to-image person search, containing 40,206 images and 80,412 textual descriptions for 13,003 unique identities. Following the official data split, the dataset is divided into a training set with 11,003 identities comprising 34,054 images and 68,108 textual descriptions; a validation set containing 1,000 identities with 3,078 images and 6,158 descriptions; and a test set also featuring 1,000 identities with 3,074 images and 6,156 descriptions. The average length of each textual description is 23 words, providing detailed visual cues for the retrieval task.

ICFG-PEDES (Ding et al., 2021b) comprises 54,522 images corresponding to 4,102 identities, with each image paired with a single textual description averaging 37 words. The training set includes 34,674 image-text pairs for 3,102 identities, while the test set consists of 19,848 image-text pairs representing the remaining 1,000 identities. This dataset is particularly notable for its one-to-one pairing of images and descriptions, emphasizing concise textual representations for each identity.

RSTPReid (Zhu et al., 2021b) contains 20,505 images from 4,101 identities captured by 15 different cameras. Each identity is represented by 5 images taken from various viewpoints, and each image is annotated with 2 textual descriptions, each containing at least 23 words. Following the standard data split, the training set consists of 3,701 identities, while the validation and test sets each contain 200 identities. The diverse camera angles and specific textual annotations make RSTPReid a valuable resource for evaluating robust retrieval methods.

B Evaluation Metrics.

To assess performance, we use the Rank-k metrics (k=1,5,10), which measure the probability of retrieving a correct match within the top-k results when queried with a textual description. In addition, we employ mean Average Precision (mAP) and mean Inverse Negative Penalty (mINP) (Ye et al., 2022), providing a more comprehensive evaluation. Higher values for Rank-k, mAP, and mINP indicate superior retrieval performance.

C Implementation Details

To evaluate the effectiveness of the proposed **CTGI** framework, we integrate it into two widely adopted TBPS baselines: **IRRA** (Jiang and Ye, 2023a) and **RDE** (Qin et al., 2024). Unless otherwise specified, we apply the same configurations and experimental protocols to both backbones to ensure fair comparison.

Backbone Architecture. Both IRRA and RDE utilize **CLIP-ViT/B-16** (Radford et al., 2021) as the image encoder and the CLIP text transformer as the text encoder. IRRA introduces an additional multimodal interaction encoder composed of transformer layers with a hidden size of 512 and 8 attention heads. Input images are resized to 384×128, and standard data augmentation is employed during training, including random horizontal flipping, cropping with padding, and random erasing.

Training Configuration. For both models, we adopt the Adam optimizer with an initial learning rate of 1×10^{-5} and a cosine decay schedule across 60 epochs. A 5-epoch linear warm-up from 1×10^{-6} is used. For randomly initialized components (e.g., IRRA's interaction encoder), a higher learning rate of 5×10^{-5} is set. The temperature parameter τ in the SDM loss is fixed at 0.02.

Extended Positional Embeddings. CLIP's default 77-token limit is insufficient for processing MTG-generated long text. Following (Zhang et al., 2024; Zhai et al., 2022), we expand the input length to **248 tokens** by retaining the first 20 learned embeddings and interpolating positions 21–77 by a factor of 4. This extension enables richer caption representations while preserving pretrained alignment.

Multimodal Language Models. The **Qwen2-VL-7B-Instruct** (Wang et al., 2024) serves as the MLLM backbone for both MTG and MTI modules, handling visual question answering and query refinement without any fine-tuning. The **OpenAI GPT-4o API** (OpenAI, 2023) is used within the Reconstructor to synthesize concise, high-quality captions from raw multi-turn QA transcripts.

Hyperparameters and Inference. During training, the MTG module performs 6 rounds of visual question-answering per image to iteratively enrich the pseudo-caption. At inference time, the MTI module conducts anchor identification by evaluating the top-K candidates (with $K = 20$) retrieved based on the initial query. Each candidate is validated via multimodal question prompts using the MLLM. If the top-ranked image surpasses a predefined similarity threshold of $\xi = 0.85$, the refinement loop may terminate early. Otherwise, the system continues checking up to 20 images and may identify multiple valid anchors (i.e., those receiving a “Yes” verdict), which are then used to jointly guide query refinement via response aggregation. The generation temperature is fixed at 0.01 to ensure output stability and reproducibility.

Hardware. All experiments are conducted on a machine equipped with two **NVIDIA GeForce RTX 4090 24GB GPUs**, providing sufficient capacity for large-scale training and inference under long-text and multi-turn interaction settings. We use mixed-precision (FP16) training to accelerate computation and reduce memory usage.

D Prompt Examples

To ensure reproducibility and offer insight into the design of our multi-turn interaction strategy, we provide representative prompts used in both **Multi-Turn Text Generation (MTG)** and **Multi-Turn Text Interaction (MTI)** modules.

D.1 Prompts for Multi-Turn Text Generation (MTG)

MTG simulates a multi-round Q&A dialogue with the MLLM to progressively enrich the visual description of a person image.

Initial Caption Prompt:

> *“Describe the person in the image as clearly and concisely as possible.”*

Refinement Questions (sampled from a predefined pool):

- *“What color is the person’s upper body clothing?”*
- *“What type of pants is the person wearing?”*
- *“Is the person carrying any objects?”*
- *“Is the person wearing any accessories (e.g., hat, bag, glasses)?”*
- *“What is the background or scene context of the image?”*
- *“Is the person performing any action?”*

Reconstruction Prompt:

> *“Rewrite the description using all the answers above, avoiding repetition while keeping it detailed and fluent.”*

D.2 Prompts for Multi-Turn Text Interaction (MTI)

During inference, MTI uses the MLLM to identify an anchor image and refine the initial user query through attribute-focused dialogue.

Anchor Verification Prompt:

> *“Does this image match the description: ‘A man in a red hoodie with black pants’? Answer yes or no.”*

Clarification Question Generation Prompt:

> *“Based on this image and the original query, suggest follow-up questions that could improve the retrieval.”*

Visual Question Answering Prompt:

> *“Please answer the following question based on the image: ‘What is the person holding?’ Answer concisely.”*

Query Aggregation Prompt:

> *“Refine the original query using the following additional details: ‘The person is wearing sunglasses and holding a white bag.’ Output a clear and discriminative new query.”*

These curated prompts guide the multi-turn reasoning process and enable CTGI to produce semantically rich training data and robust test-time refinements. Additional prompt sets and template variations are provided in our released code repository.