

Anonymous Authors

ROGLE: Robust Global-Local Alignment with Automated Region Supervision for Text-Based Person Search

Chaodong Jia

Chengdu University of Information Technology, Chengdu, China
jingoja505@gmail.com

Abstract. Text-Based Person Search (TBPS) focuses on identifying individuals from a gallery solely based on descriptive text, without relying on example images. However, existing TBPS models, especially those based on CLIP, struggle with fine-grained understanding due to global representational bias and semantic sparsity inherited from training on short captions. This results in weak fine-grained alignment, exacerbated by the scarcity of region-level annotations. To address this, we propose ROGLE (Robust Global-Local Embedding), a unified framework that overcomes reliance on costly manual annotations through an automated Region-to-Sentence Matching (RSM) strategy. RSM automatically mines pseudo region-sentence pairs for scalable fine-grained supervision. Furthermore, ROGLE employs a multi-granular learning strategy that fuses global contrastive learning with region-level local alignment. We also introduce the P-VLG Benchmark, a large-scale dataset constructed by curating and enriching images from established public benchmarks. It features over 100,000 annotated regions and rich long-form captions, making it the first TBPS benchmark to support both global and local assessment protocols. Extensive experiments show that ROGLE significantly outperforms existing approaches, particularly on challenging long-form queries. Code and the P-VLG benchmark will be made publicly available.

Keywords: P-VLG Benchmark, fine-grained alignment, Text-Based Person Search

1 Introduction

Text-Based Person Search (TBPS) seeks to find images of individuals in large galleries using only textual descriptions, which is crucial for scenarios like surveillance and missing-person search. Unlike existing person re-ID strategies that need sample images for comparison, TBPS better models practical scenarios such as surveillance analysis or missing-person investigations, where no visual examples are available and verbal witness reports serve as the only source of information. Consequently, TBPS plays a vital role in enabling retrieval systems to operate effectively in open-world and zero-shot settings.

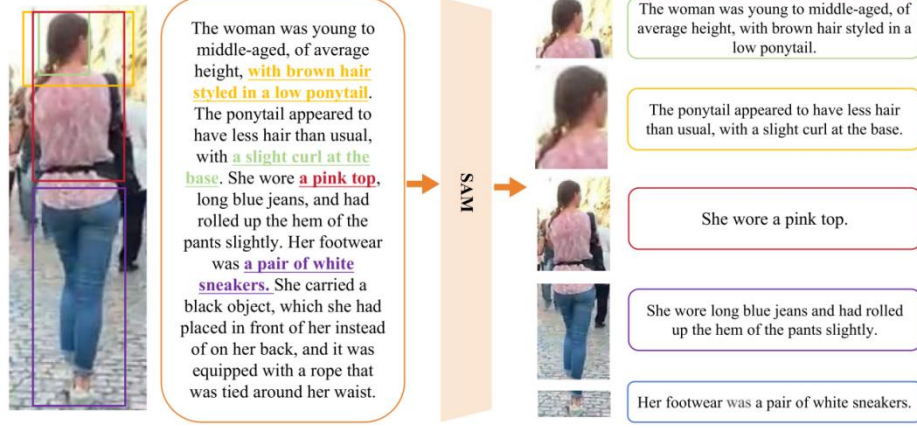


Fig. 1. Visualization of our Region-to-Sentence Matching (RSM) pipeline. The original person image (left) is paired with a long descriptive caption that is first decomposed into independent sentences. We apply SAM to segment the image into semantic regions (colored boxes) and compute cross-modal similarity between each region and each sentence. The highest-scoring sentence for each region is selected, producing five local region-sentence pairs on the right, corresponding to hairstyle, clothing, and footwear.

TBPS research has benefited from datasets including CUHK-PEDES, ICFG-PEDES, and RSTPReid. However, many current approaches, which rely on large pre-trained vision-language models like CLIP, still face challenges in capturing fine-grained details. These models, trained on brief and generic captions, often emphasize global image semantics while overlooking localized features. Consequently, training and evaluation datasets contain mainly brief, attribute-based annotations, offering only coarse-grained supervision that restricts a model’s ability to distinguish visually similar individuals or interpret complex queries. Moreover, natural language descriptions often include abstract relations or localized cues that do not directly correspond to visual regions, making fine-grained alignment challenging. The lack of region-level annotations further constrains learning meaningful part-level correspondences.

To address the data-level bottleneck, we construct a large-scale benchmark named P-VLG using a novel Region-to-Sentence Matching (RSM) module. RSM decomposes long-form textual descriptions into atomic sentences and aligns them with semantically coherent image regions extracted by SAM. By leveraging CLIP embeddings to compute cross-modal similarity, RSM automatically mines pseudo region-sentence pairs, providing fine-grained supervision without manual annotation. The P-VLG benchmark overcomes the limitations of prior TBPS datasets, which often lack dense grounding and favor short captions. It contains 6,801 training identities, 48,485 images, and 68,990 enriched captions, with a test set of 400 unseen identities and 6,658 image-caption pairs. The validation set provides over 100,000 region-caption alignments, enabling evaluation of both global retrieval and region-level grounding.

Based on these pseudo-labels, ROGLE adopts a dual-branch training paradigm. The first branch performs global contrastive learning to align whole-image and full-caption embeddings, while the second branch provides fine-grained supervision through multi-granular learning that integrates global and region-level local alignment. This design encourages the model to learn both contextualized global semantics and discriminative local features, essential for resolving ambiguous or compositional queries. In addition, a reliability-aware alignment mechanism enhances robustness against noisy correspondences in large-scale data. Our contributions are summarized as follows:

- We propose ROGLE, a unified global–local framework tackling semantic sparsity and misalignment. It incorporates an automated Region-to-Sentence Matching (RSM) module to mine pseudo-alignments from long captions, enabling fine-grained supervision without manual annotation.
- We introduce the P-VLG Benchmark, the first dataset supporting multi-granular TBPS evaluation. It features enriched descriptions and over 100,000 region-level annotations to assess both holistic and localized alignment.
- We design a multi-granular learning strategy synergizing global contrastive learning with region-level alignment. Extensive experiments demonstrate superior performance on challenging benchmarks, particularly for detailed and compositional queries.

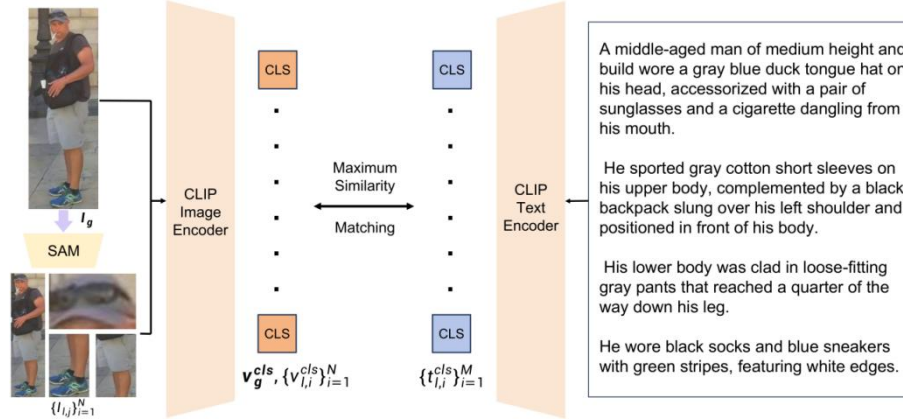


Fig. 2. Visualization of the Region-to-Sentence Matching process in our P-VLG Benchmark. Given a pedestrian image and its long-form caption, SAM is used to generate region proposals from the global image. All image regions and textual sentences are encoded with CLIP, and cross-modal similarity is computed for all region-sentence pairs. The highest-scoring pair is then selected to create pseudo-aligned fine-grained supervision.

2 Related Work

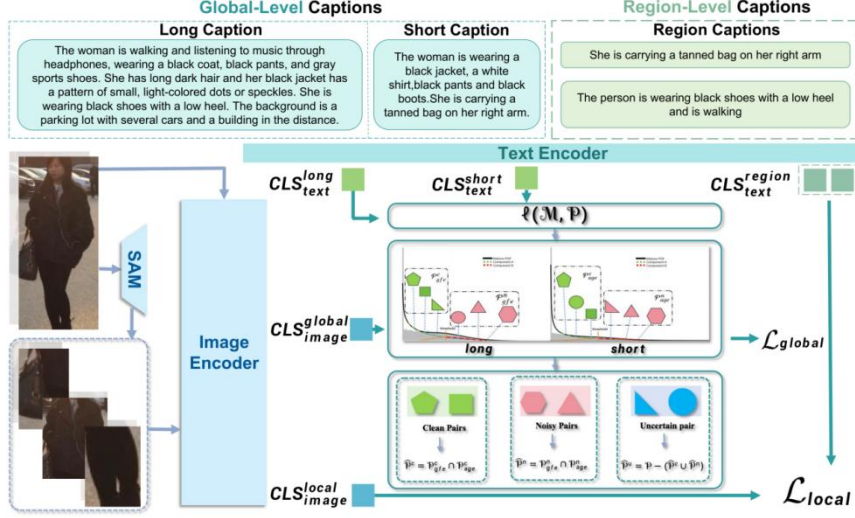


Fig. 3. Overview of ROGLe. SAM segments the input image into local regions. We employ three textual supervisions: global long, concise short, and region-level captions. Dual encoders extract text () and image () representations. A probabilistic module estimates pair reliability (clean/noisy/uncertain) to guide global () and local () alignment.

2.1 Text-Based Person Search

TBPS focuses on identifying individuals within extensive image datasets solely from descriptive text. Traditional approaches typically employ global contrastive learning to align full-image and text representations. Although effective for coarse semantic matching, these methods often fail to capture fine-grained visual cues, including accessories, hairstyles, or layered clothing, which are essential for distinguishing similar identities. This limitation motivates approaches that integrate both holistic and fine-grained alignment.

2.2 Fine-Grained Alignment

To achieve finer alignment, existing strategies fall into two categories. Manual or part-based localization methods align phrases with corresponding image regions but rely heavily on labor-intensive annotations or specialized part-parsing modules, limiting scalability and generalization. Detector-based supervision approaches, such as ViTAA, CLOC, and OWLv2, perform localized alignment via attribute modeling or open-vocabulary detection. While these methods achieve strong spatial correspondence, they often require large-scale pretraining and high computational cost. To overcome these challenges, we introduce the automatic Region-to-Sentence Matching (RSM) mechanism, which leverages pretrained CLIP features and

segmentation masks from SAM to generate high-quality local supervision without manual annotations. This annotation-free, global-local alignment framework supports scalable fine-tuning while maintaining strong performance on TBPS tasks.

3 Method

We present ROGLE, a unified framework designed to enhance fine-grained understanding in TBPS through a global-local learning paradigm. Unlike prior methods that handle global and local representations separately or assume perfect modality alignment, ROGLE integrates scalable pseudo-pair generation with hierarchical training. This design leverages both the descriptive richness of textual annotations and the structural complexity of pedestrian images, enabling robust alignment at multiple granularities.

3.1 Problem Statement

In TBPS, the goal is to identify target individuals within an image collection using only textual descriptions. Let the gallery be $V = \{I_i, y_i^p, y_i^v\}_{i=1}^{N_v}$, where I_i denotes the i -th image, $y_i^p \in \{1, \dots, C\}$ is the person identity, and $y_i^v \in \{1, \dots, N_v\}$ is the image-specific index. The corresponding textual description set is $T = \{T_j, y_j^v\}_{j=1}^{N_t}$, where T_j describes the pedestrian in image I_j .

During training, the dataset consists of cross-modal pairs $P = \{(I_i, T_j), l_{ij}\}_{i=1}^N$, where $l_{ij} \in \{0, 1\}$ indicates whether the image and text correspond to the same person. A positive pair ideally satisfies both identity and image alignment, i.e., $y_i^p = y_j^p$ and $y_i^v = y_j^v$. However, real-world datasets often contain noisy correspondences due to annotation errors, which can introduce false positives and degrade retrieval performance. TBPS therefore requires models capable of handling both global semantic alignment and fine-grained visual distinctions, such as clothing details, accessories, and localized attributes, while being robust to noisy supervision.

3.2 Generating the P-VLG Benchmark

Region-to-Sentence Matching

To enable fine-grained supervision without manual annotation, we introduce the **Region-to-Sentence Matching (RSM)** module. Given a pedestrian image I_g and its associated long-form caption T_g , RSM aims to generate pseudo region-sentence pairs (I_l, T_l) that capture localized visual-text correspondences. The process consists of two main steps:

- 1. Sentence decomposition:** The long-form caption T_g is segmented into M atomic sentences $\{T_l^{(j)}\}_{j=1}^M$ using the spaCy library. Each sentence typically describes a specific visual attribute, such as clothing, hairstyle, or accessories.

2. Region localization: We apply the Segment Anything Model (SAM) with a ViT-H backbone in automatic mode to obtain N semantic masks $\{M^{(i)}\}_{i=1}^N$ from I_g . Each mask is converted into a bounding box $B^{(i)}$ to crop local image regions $\{I_l^{(i)}\}_{i=1}^N$. Trivial regions occupying less than 1% of the image are discarded, typically resulting in 2–5 valid masks per image.

Each image region and sentence is processed through a pre-trained CLIP (ViT-B/16) encoder to extract visual and textual features:

$$\mathbf{t}_l^{(i)} = f^I(I_l^{(i)}), \quad \mathbf{t}_l^{(j)} = f^T(T_l^{(j)}),$$

We then compute the cosine similarity between all region–sentence combinations and select the pair with the highest score:

$$(i^*, j^*) = \underset{i,j}{\operatorname{argmax}} \cos(\mathbf{t}_l^{(i)}, \mathbf{t}_l^{(j)}).$$

This process uses a greedy assignment strategy to ensure a one-to-one alignment between regions and sentences. If the selected region $I_l^{(i^*)}$ covers nearly the entire image, it is discarded to maintain locality. Otherwise, the aligned pseudo pair $(I_l, T_l) = (I_l^{(i^*)}, T_l^{(j^*)})$ is used for fine-grained supervision. This automatic procedure allows the model to leverage localized visual cues without relying on labor-intensive manual annotations.

P-VLG Benchmark

Construction and Data Provenance. To facilitate large-scale fine-grained learning while adhering to ethical data standards, we introduce the **P-VLG Benchmark**. The benchmark dataset integrates images from CUHK-PEDES, ICFG-PEDES, RSTPReid, and UfineBench, ensuring compliance with privacy standards and enabling large-scale fine-grained learning. By leveraging these existing resources, we ensure compliance with privacy protocols while introducing rich, multi-granular annotations that overcome the semantic sparsity of original short captions.

The benchmark contains 6,801 training identities with 48,485 images and 68,990 enriched long-form captions. The validation set includes over 100,000 region–caption alignments, providing a valuable resource for evaluating local grounding. The test set consists of 400 unseen identities with 6,658 image–caption pairs, supporting both global and region-level retrieval evaluation.

Annotation Process. All region–caption pairs are automatically generated using the Region-to-Sentence Matching (RSM) module. In this process, long-form captions are decomposed into atomic sentences, while images are segmented into semantic regions using SAM. CLIP features are extracted for both sentences and regions, and the most similar sentence–region pairs are selected to form pseudo region–sentence alignments. This automated approach provides dense, high-quality supervision without manual annotation, making P-VLG the first TBPS benchmark to support a multi-granular evaluation protocol that jointly assesses global and local alignment performance.

3.3 The ROGLE Framework

ROGLE synergizes coarse and fine-grained learning via three components: a dual-encoder backbone, a reliability-guided mechanism for noise robustness, and a collaborative objective for multi-granular alignment.

Dual-Encoder Feature Extraction

We employ a dual-encoder setup, with separate encoders for images and text, both initialized with CLIP pre-trained weights to enable effective cross-modal feature alignment.

Image Encoder

The image encoder is based on a Vision Transformer (ViT). Input images are divided into non-overlapping patches, which are then converted into token embeddings. A learnable [CLS] token is added at the beginning of the sequence. The token sequence passes through multiple Transformer layers, and the [CLS] token output is projected into the joint embedding space to obtain the global image representation:

$$\mathbf{f}^v = f^v(I).$$

Text Encoder

For textual input, we employ a Transformer-based text encoder. The input is tokenized using BPE, and special tokens [SOS] and [EOS] frame the sequence. After processing through Transformer layers, the hidden state corresponding to the [EOS] token is mapped to the joint embedding space, producing the global text embedding:

$$\mathbf{f}^t = f^t(T).$$

Reliability-Guided Alignment

Real-world datasets often contain noisy correspondences. To address this, we adopt a reliability-guided calibration mechanism inspired by RDE to dynamically handle uncertain training pairs.

During training, each pair (I_i, T_i) in a batch of size N_b is first used to compute its alignment loss $l_i = L(I_i, T_i)$. We apply a two-component Gaussian Mixture Model (GMM) on the loss distributions for both long and short captions to determine the likelihood that a pair is reliable ($k=1$) or unreliable ($k=0$). Based on these probabilities, the pairs are categorized into three groups: clean ($\hat{P}^c$), noisy ($\hat{P}^n$), and uncertain ($\hat{P}^u$). Each sample is then assigned a reliability-aware weight $\hat{w}_i$ to prioritize learning from high-quality pairs while reducing the impact of noisy ones:

$$\hat{w}_i = \begin{cases} 2, & \text{if } (I_i, T_i) \in \hat{P}^c, \\ 0, & \text{if } (I_i, T_i) \in \hat{P}^n, \\ 1, & \text{otherwise.} \end{cases}$$

This allows the model to prioritize learning from clean pairs while reducing the influence of noisy ones.

Global-Local Collaborative Learning

ROGLE learns features at multiple granularities by jointly supervising coarse and fine-grained alignment. Given global pairs (I_g, T_g) and local pseudo-pairs (I_l, T_l) from our RSM module, the encoder extracts their respective embeddings:

$$\begin{aligned} \mathbf{f}_g^v &= f^v(I_g), & \mathbf{f}_g^t &= f^t(T_g) \\ \mathbf{f}_l^v &= f^v(I_l), & \mathbf{f}_l^t &= f^t(T_l) \end{aligned}$$

The learning is driven by a dual-branch loss: **Global Loss** (L_{global}): This loss facilitates coarse-grained alignment between the global embeddings of entire images and captions. Following IRRA, it is defined as:

$$L_{\text{global}} = L_{\text{SDM}} + L_{\text{InfoNCE}}$$

Local Loss (L_{local}): This loss ensures fine-grained alignment between local image regions and sentence fragments. It is a contrastive loss over the local embeddings in a batch:

$$L_{\text{local}} = -\frac{1}{N_b} \sum_{n=1}^{N_b} \log \frac{\exp(\cos(\mathbf{f}_l^{v(n)}, \mathbf{f}_l^{t(n)})/\tau)}{\sum_{k=1}^{N_b} \exp(\cos(\mathbf{f}_l^{v(n)}, \mathbf{f}_l^{t(k)})/\tau)}.$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity, and τ serves as the temperature parameter.

Training Objective

The total training objective is defined by integrating both global and local loss components, each scaled according to its reliability weight, forming a robust framework that supervises multiple granularities. The total loss is:

$$L_{\text{total}} = \frac{1}{N_b} \sum_{i=1}^{N_b} \hat{w}_i L_{\text{global}}^{(i)} + L_{\text{local}}$$

For the i -th sample, $L_{\text{global}}^{(i)}$ represents the global loss. The two-branch supervision mechanism allows ROGLE to align holistic image-text semantics as well as detailed local features, enabling accurate and robust retrieval of images based on textual descriptions. The model's performance is evaluated on the CUHK-PEDES dataset, with the highest results highlighted in the tables.

Table 1. Performance comparisons on the CUHK-PEDES dataset. The best results are in bold.

Methods	Ref.	Image Enc.	Text Enc.	R-1	R-5	R-10	mAP	mNP
ViTAA	ECCV'20	RN50	LSTM	54.92	75.18	82.90	51.60	-
DSSL	ACMMM'21	RN50	BERT	59.98	80.41	87.56	-	-
SSAN	arXiv'21	RN50	LSTM	61.37	80.15	86.73	-	-
LapScore	ICCV'21	RN50	BERT	63.40	-	87.80	-	-
LBUL	ACMMM'22	RN50	BERT	64.04	82.66	87.22	-	-
LGUR	ACMMM'22	DeiT-Small	BERT	65.25	83.12	89.00	-	-
DCEL	ACMMM'22	CLIP-ViT	CLIP-Xformer	71.36	88.11	92.48	64.25	48.26
IVT	ECCV'22	ViT-Base	BERT	65.59	83.11	89.21	-	-
CFine	TIP'23	CLIP-ViT	BERT	69.57	85.93	91.15	-	-
IRRA	CVPR'23	CLIP-ViT	CLIP-Xformer	73.38	89.93	93.71	66.13	50.24
BiLMa	ICCV'23	CLIP-ViT	CLIP-Xformer	74.03	89.59	93.62	66.57	-

Methods	Ref.	Image Enc.	Text Enc.	R-1	R-5	R-10	mAP	mINP
DCEL	ACMMM'23	CLIP-ViT	CLIP-Xformer	75.02	90.89	94.52	-	-
APTM	ACMMM'23	Swin-Transformer	Bert	76.53	90.04	94.15	66.91	-
RaSa	IJCAI'23	ALBEFF	ALBEFF	76.51	90.29	94.25	69.38	-
RDE	CVPR'24	CLIP-ViT	CLIP-Xformer	75.94	90.14	94.12	67.56	51.44
CFAM	CVPR'24	CLIP-ViT	CLIP-Xformer	75.60	90.53	-	67.27	-
PLOT	ECCV'24	CLIP-ViT	CLIP-Xformer	75.28	90.42	94.12	-	-
DM-Adapter	AAAI'25	CLIP-ViT	CLIP-Xformer	72.17	88.74	92.85	64.33	-
OCDL	ICASSP'25	CLIP-ViT	CLIP-Xformer	75.10	89.43	-	68.18	-
ROGLE (Ours)	-	CLIP-ViT	CLIP-Xformer	76.93	91.24	95.00	69.14	53.23

Table 2. Performance comparisons on the ICFG-PEDES dataset. The best results are in bold.

Methods	R-1	R-5	R-10	mAP	mINP
Dual Path	38.99	59.44	68.41	-	-
ViTAA	50.98	68.79	75.78	-	-
SSAN	54.23	72.63	79.53	-	-
IVT	56.04	73.60	80.22	-	-
CFine	60.83	76.55	82.42	-	-
IRRA	63.46	80.25	85.82	38.06	7.93
BiLMa	63.83	80.15	85.74	38.26	-
DCEL	64.88	81.34	86.72	-	-
RaSa	65.28	80.40	85.12	41.29	-
CFAM	65.38	81.17	-	39.42	-
PLOT	65.76	81.39	86.73	-	-
ROGLE (Ours)	66.97	82.27	87.21	39.81	7.88

Table 3. Performance Comparison on the RSTPReid Dataset. The best results are in bold.

Methods	R-1	R-5	R-10	mAP	mINP
DSSL	39.05	62.60	73.95	-	-
SSAN	43.50	67.80	77.15	-	-
IVT	46.70	70.00	78.80	-	-
CFine	50.55	72.50	81.60	-	-
IRRA	60.20	81.30	88.20	47.17	25.28
DCEL	61.35	83.95	90.45	-	-

Methods	R-1	R-5	R-10	mAP	mINP
RaSa	66.90	86.50	91.35	52.31	-
PLOT	61.80	82.85	89.45	-	-
ROGLE (Ours)	65.20	83.75	89.45	50.93	27.83

Table 4. Performance comparisons on the P-VLG dataset. The best results are in bold.

Methods	R-1	R-5	R-10	mAP	mINP
DECL	73.78	86.66	90.80	53.09	19.72
IRRA	73.30	86.57	90.95	54.12	21.22
RDE	68.82	83.66	88.72	48.10	15.75
ROGLE (Ours)	78.03	89.17	92.60	57.82	24.01

Table 5. Ablation study of ROGLE components on CUHK-PEDES. The best results are in bold.

Method	L_{global}	L_{local}	RGAP	R-1	R-5	mAP
Global only				71.56	88.23	64.38
Local only				69.92	87.15	63.12
Global + Local				74.87	90.02	67.25
ROGLE (full)				76.93	91.24	69.14

Table 6. Cross-domain evaluation. P-VLG trained, tested on other datasets.

Method	Domain	R1	R5	mAP
ROGLE	P-VLG $\rightarrow$ ICFG	62.78	78.19	38.58
ROGLE	P-VLG $\rightarrow$ CUHK	47.74	68.65	43.10

4 Experiments

This section reports comprehensive experiments evaluating the performance and benefits of the ROGLE framework.

4.1 Datasets and Settings

The CUHK-PEDES dataset, commonly used for text-to-image person retrieval, comprises 40,206 images and 80,412 textual descriptions corresponding to 13,003 individuals. Among them, 11,003 identities are designated for training, while 1,000 identities each are reserved for validation and testing.

ICFG-PEDES contains 54,522 image-text pairs from 4,102 identities, with 3,102 for training and 1,000 for testing.

RSTPReid includes 20,505 images of 4,101 identities captured across 15 cameras. Each image is annotated with two textual descriptions. The dataset is partitioned into 3,701 identities for training, 200 for validation, and 200 for testing.

4.2 Implementation Details

ROGLE leverages CLIP with a ViT-B/16 backbone for image encoding and a Transformer-based text encoder, both initialized with the original 400M pre-trained weights. Region are generated by SAM (ViT-H) in automatic mode, producing 2–5 valid masks per image after filtering trivial regions. Captions are decomposed into short atomic sentences using spaCy, and region–sentence matching is computed via cosine similarity with greedy assignment to ensure one-to-one alignment. Images are augmented by random flip, crop, erase, and color jitter, then resized to 384×128 . Texts are tokenized by CLIP’s BPE (max length 77). The model is trained for 60 epochs using the Adam optimizer with a batch size of 64 and a weight decay of 5×10^{-4} . The learning rate follows a two-stage schedule, starting with a linear warm-up over 5 epochs, followed by cosine decay. The contrastive temperature τ is 0.02. The reliability calibration updates a GMM every 5 epochs (from epoch 10) with threshold 0.5. All experiments run on a single RTX 4090 GPU. Training takes 8–10 hours per dataset, and inference speed reaches 128 image–text pairs per second. The model has 158M parameters.

4.3 Evaluation Protocols

We use standard Rank-K metrics ($K = 1, 5, 10$) to assess retrieval performance. In addition, we report mean Average Precision (mAP) and mean Inverse Negative Penalty (mINP) to provide a thorough assessment of the model’s performance.

4.4 Comparison with State-of-the-Art

On the CUHK-PEDES dataset, ROGLE attains a Rank-1 accuracy of 76.93%, as reported in Table 1, exceeding previously published methods. These results demonstrate the benefit of combining global and local supervision during training. For the ICFG-PEDES dataset, ROGLE achieves 66.97% Rank-1 accuracy (Table 2), outperforming earlier approaches such as IRRA (63.46%) by a notable margin. Improvements are also observed in mAP and mINP, indicating the method’s robustness.

On RSTPReid, the model reaches 65.20% Rank-1 accuracy (Table 3), considerably higher than IVT (51.7%) and Cfine (55.55%). This confirms ROGLE’s capability to manage high-resolution images and complex textual descriptions in dense urban settings..

4.5 Ablation Study

We validate ROGLe on CUHK-PEDES (Table 5). The global-only baseline achieves 71.56% Rank-1, while local-only alignment yields 69.92%, indicating neither is sufficient alone. Combining both strategies significantly improves accuracy to 74.87% (+3.31%), demonstrating their complementary nature. Furthermore, adding the RGAP module boosts Rank-1 to 76.93% (+2.06%) by adaptively filtering noisy pseudo-labels via Beta Mixture Modeling. These results confirm that effective TBPS requires synergizing global coherence, local discriminative details, and noise-robust training.

5 Conclusion

This work introduces ROGLe, a unified framework for fine-grained text-to-image person retrieval, which leverages automatic region-sentence alignment and multi-granular learning. To address the scarcity of fine-grained supervision, we incorporate the RSM module that generates reliable pseudo region-sentence pairs from long-form captions, enabling effective multi-level supervision. Additionally, we present the P-VLG Benchmark, a large-scale dataset with over 100,000 region-level annotations and long-form captions, supporting both global and local evaluation protocols. Experimental results demonstrate that ROGLe consistently achieves competitive performance across multiple challenging datasets, including CUHK-PEDES, ICFG-PEDES, RSTPReid, and P-VLG. These findings confirm the effectiveness, robustness, and practical applicability of our approach in real-world scenarios.

6 References

1. Li, Shuang, Xiao, Tong, Li, Hongsheng, et al.: Person search with natural language description. In: CVPR, pp. 1970-1979 (2017)}
2. Ding, Zefeng, Ding, Changxing, Shao, Zhiyin, et al.: Semantically self-aligned network for text-to-image part-aware person re-identification. arXiv preprint arXiv:2107.12666 (2021)}
3. Zhu, Aichun, Wang, Zijie, Li, Yifeng, et al.: Dssl: Deep surroundings-person separation learning for text-based person retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia, pp. 209-217 (2021)}
4. Radford, Alec, Kim, Jong Wook, Hallacy, Chris, et al.: Learning transferable visual models from natural language supervision. In: ICML, pp. 8748-8763 (2021)
5. Kirillov, Alexander, Mintun, Eric, Ravi, Nikhila, et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 4015-4026 (2023)}
6. Honnibal, Matthew, Montani, Ines: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, vol. 7, pp. 411-420 (2017)}
7. Kingma, Diederik P, Ba, Jimmy: Adam: A Method for Stochastic Optimization. In: ICLR (Poster) (2015)}

8. Jiang, Ding, Ye, Mang: Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. In: CVPR, pp. 2787-2797 (2023)}
9. Wang, Zhe, Fang, Zhiyuan, Wang, Jun, et al.: Vitaa: Visual-textual attributes alignment in person search by natural language. In: Computer Vision--ECCV 2020: 16th European Conference, Glasgow, UK, August 23--28, 2020, Proceedings, Part XII 16, pp. 402-420 (2020)
10. Zhang, Ying, Lu, Huchuan: Deep cross-modal projection learning for image-text matching. In: Proceedings of the European conference on computer vision (ECCV), pp. 686-701 (2018)}
11. Wu, Yushuang, Yan, Zizheng, Han, Xiaoguang, et al.: LapsCore: language-guided person search via color reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1624-1633 (2021)}
12. Wang, Zijie, Zhu, Aichun, Xue, Jingyi, et al.: Look Before You Leap: Improving Text-based Person Retrieval by Learning A Consistent Cross-modal Common Manifold. In: ACM MM, pp. 1984-1992 (2022)}
13. Shao, Zhiyin, Zhang, Xinyu, Fang, Meng, et al.: Learning Granularity-Unified Representations for Text-to-Image Person Re-identification. In: ACM MM, pp. 5566-5574 (2022)}
14. Qin, Yang, Peng, Dezhong, Peng, Xi, et al.: Deep evidential learning with noisy correspondence for cross-modal retrieval. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4948-4956 (2022)}
15. Shu, Xiujun, Wen, Wei, Wu, Haoqian, et al.: See finer, see more: Implicit modality alignment for text-based person retrieval. In: European Conference on Computer Vision, pp. 624-641 (2022)
16. Yan, Shuanglin, Dong, Neng, Zhang, Liyan, et al.: Clip-driven fine-grained text-image person re-identification. IEEE Transactions on Image Processing (2023)
17. Fujii, Takuro, Tarashima, Shuhei: BiLMA: Bidirectional Local-Matching for Text-based Person Re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2786-2790 (2023)}
18. Shen, Fei, Shu, Xiangbo, Du, Xiaoyu, et al.: Pedestrian-specific bipartite-aware similarity learning for text-based person retrieval. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 8922-8931 (2023)}
19. Ma, Yiwei, Sun, Xiaoshuai, Ji, Jiayi, et al.: Beat: Bi-directional One-to-Many Embedding Alignment for Text-based Person Retrieval. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 4157-4168 (2023)}
20. Yan, Shuanglin, Dong, Neng, Liu, Jun, et al.: Learning comprehensive representations with richer self for text-to-image person re-identification. In: Proceedings of the 31st ACM international conference on multimedia, pp. 6202-6211 (2023)}
21. Li, Shenshen, Xu, Xing, Yang, Yang, et al.: DCEL: Deep Cross-modal Evidential Learning for Text-Based Person Retrieval. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 6292-6300 (2023)}
22. Shao, Zhiyin, Zhang, Xinyu, Ding, Changxing, et al.: Unified pre-training with pseudo texts for text-to-image person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11174-11184 (2023)}
23. Bai, Yang, Cao, Min, Gao, Daming, et al.: RaSa: relation and sensitivity aware representation learning for text-based person search. In: IJCAI, pp. 555-563 (2023)}
24. Yang, Shuyu, Zhou, Yinan, Wang, Yaxiong, et al.: Towards Unified Text-based Person Retrieval: A Large-scale Multi-Attribute and Language Search Benchmark. In: Proceedings of the 2023 ACM on Multimedia Conference (2023)}
25. Farooq, Ammarah, Awais, Muhammad, Kittler, Josef, et al.: AXM-Net: Implicit Cross-Modal Feature Alignment for Person Re-identification. In: AAAI, pp. 4477-4485 (2022)}

26. Wang, Zijie, Zhu, Aichun, Xue, Jingyi, et al.: CAIBC: Capturing All-round Information Beyond Color for Text-based Person Retrieval. In: ACM MM, pp. 5314-5322 (2022)}
27. Cao, Min, Bai, Yang, Zeng, Ziyin, et al.: An Empirical Study of CLIP for Text-Based Person Search. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 465-473 (2024)}
28. Song, Zifan, Hu, Guosheng, Zhao, Cairong: Diverse person: Customize your own dataset for text-based person search. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 4943-4951 (2024)}
29. Liu, Yu, Qin, Guihe, Chen, Haipeng, et al.: Causality-Inspired Invariant Representation Learning for Text-Based Person Retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 14052-14060 (2024)}
30. Zhao, Zhiwei, Liu, Bin, Lu, Yan, et al.: Unifying Multi-Modal Uncertainty Modeling and Semantic Alignment for Text-to-Image Person Re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 7534-7542 (2024)}
31. Qin, Yang, Chen, Yingke, Peng, Dezhong, et al.: Noisy-Correspondence Learning for Text-to-Image Person Re-identification. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
32. Zuo, Jialong, Zhou, Hanyu, Nie, Ying, et al.: Ufinebench: Towards text-based person retrieval with ultra-fine granularity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22010-22019 (2024)}
33. Jicheol Park, Dongwon Kim, Boseung Jeong, et al.: PLOT: Text-based Person Search with Part Slot Attention for Corresponding Part Discovery. (2024)
34. Lv, Tianle, Li, Shuang, Leng, Jiaxu, et al.: 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), pp. 2895-2899 (2024)
35. Liu, Yating, Liu, Zimo, Lan, Xiangyuan, et al.: Dm-adapter: Domain-aware mixture-of-adapters for text-based person retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 5703-5711 (2025)}
36. Li, Haiwen, Liu, Delong, Su, Fei, et al.: Object-Centric Discriminative Learning for Text-Based Person Retrieval. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1-5 (2025)
37. Sarafianos, Nikolaos, Xu, Xiang, Kakadiaris, Ioannis A: Adversarial representation learning for text-to-image matching. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 5814-5824 (2019)}
38. Han, Xiao, He, Sen, Zhang, Li, et al.: Text-based person search with limited data. arXiv:2110.10807 (2021)}
39. Li, Shiping, Cao, Min, Zhang, Min: Learning semantic-aligned feature representation for text-based person search. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2724-2728 (2022)
40. Zheng, Zhedong, Zheng, Liang, Garrett, Michael, et al.: Dual-path convolutional image-text embeddings with instance loss. ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 16, pp. 1-23 (2020)
41. Qin, Yaqing, Peng, Donghao, Peng, Xi, et al.: Deep Evidential Learning with Noisy Correspondence for Cross-Modal Retrieval. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 4948-4956 (2022)
42. Yan, Shuanglin, Tang, Hao, Zhang, Liyan, et al.: Image-specific information suppression and implicit local alignment for text-based person search. arXiv preprint arXiv:2208.14365 (2022)}
43. Wang, Chengji, Luo, Zhiming, Lin, Yaojin, et al.: Text-based Person Search via Multi-Granularity Embedding Learning. In: IJCAI, pp. 1068-1074 (2021)}

44. Luo, Hao, Gu, Youzhi, Liao, Xingyu, et al.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 0-0 (2019)}
45. He, Shuting, Luo, Hao, Wang, Pichao, et al.: Transreid: Transformer-based object re-identification. In: ICCV, pp. 15013-15022 (2021)}
46. Gao, Chenyang, Cai, Guanyu, Jiang, Xinyang, et al.: Contextual Non-Local Alignment over Full-Scale Representation for Text-Based Person Search. arXiv preprint arXiv:2101.03036 (2021)}
47. Liu, Delong, Li, Haiwen, Zhao, Zhicheng, et al.: Text-guided image restoration and semantic enhancement for text-to-image person retrieval. Neural Networks, vol. 184, pp. 107028 (2025)
48. Xie, Zequn, Ji, Haoming, Meng, Lingwei: Dynamic Uncertainty Learning with Noisy Correspondence for Text-Based Person Search. arXiv preprint arXiv:2505.06566 (2025)}
49. Minderer, Matthias, Gritsenko, Alexey, Hounsby, Neil: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems, vol. 36, pp. 72983-73007 (2023)}
50. Chen, Hong-You, Lai, Zhengfeng, Zhang, Haotian, et al.: Contrastive localized language-image pre-training. arXiv preprint arXiv:2410.02746 (2024)}
51. Radford, Alec, Kim, Jong Wook, Hallacy, Chris, et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp. 8748-8763 (2021)
52. Kirillov, Alexander, Mintun, Eric, Ravi, Nikhila, et al.: Segment Anything. arXiv:2304.02643 (2023)}