International Journal of Multimedia Information Retrieval (2025) 14:11
https://doi.org/10.1007/s13735-025-00356-w

REGULAR PAPER

Cross-modal alignment with synthetic caption for text-based person search

Weichen Zhao[1,3] · Yuxing Lu[2] · Zhiyuan Liu[3] · Yuan Yang[1] · Ge Jiao[1]

Received: 28 August 2024 / Revised: 30 November 2024 / Accepted: 12 February 2025 / Published online: 10 March 2025

Abstract

Text-based person search aims to retrieve target person from a large gallery based on natural language description. Existing methods take it as one-to-one embedding or many-to-many embedding matching problem. The former approach relies on the assumption of the existence of strong alignment between text and images, while the latter inevitably leads to issues of intra-class variation. Rather than being confined to these two approaches, we propose a new strategy that achieves cross-modal alignment with synthetic caption for joint image-text-caption optimization, named CASC. The core of this strategy lies in generating fine-grained captions that are informative for multimodal alignment. To realize this, we introduce two novel components: Granularity Awareness Sensor (GAS) and Conditional Contrastive Learning (CCL). GAS selects relative features through an innovative adaptive masking strategy, endowing the model with an enhanced perception of discriminative features. CCL aligns different modalities through further constraints on the synthetic captions by comparing the similarity of hard negative samples, protecting the disruption from noisy contents. With the incorporation of extra caption supervision, the model has access to learn more comprehensive feature representation, which in turn boosts the retrieval performance during inference. Experiments demonstrate that CASC outperforms existing state-of-the-art methods by 1.20%, 2.35% and 2.29% in terms of Rank@1 on CUHK-PEDES, ICFG-PEDES and RSTPReid datasets, respectively.

Keywords Text-based person search · Cross-modal retrieval · Cross-modal alignment · Synthetic caption

1 Introduction

Recently, Text-based person search (TBPS) has gained considerable popularity in cross-modal retrieval. It seeks to locate a specific person in a gallery of images by using a descriptive text query. Despite its user-friendly and flexible querying process using natural language, this approach comes with its own complexities. The primary challenge in TBPS lies in bridging the modality gap between textual queries and visual data, caused by inherent differences in their semantic representations. This gap complicates the alignment and comparison of cross-modal features. Furthermore, TBPS suffers from limited training data [1], as large-scale datasets comparable to those in Image-Text Retrieval (ITR) are challenging to collect due to privacy concerns and the time-intensive nature of text annotation. These challenges necessitate innovative solutions specifically tailored to TBPS, extending beyond the capabilities of existing ITR methods.

To address the modality gap, prior works [2–4] often adopt a one-to-one embedding framework (as shown in Fig. 1a),

✉ Ge Jiao
jiaoge@126.com

Weichen Zhao
wczhao2229@163.com

Yuxing Lu
luyx@stu.pku.edu.cn

Zhiyuan Liu
20224227015@stu.suda.edu.cn

Yuan Yang
yang_yuan2023@163.com

1 College of Computer Science and Technology, Hengyang Normal University, Henghua Rd., Zhuhui District, Hengyang 421010, China

2 College of Future Technology, Peking University, Yiheyuan Rd., Haitian District, Beijing 100871, China

3 School of Computer Science and Technology, Soochow University, Shizi Street, Suzhou 215000, China

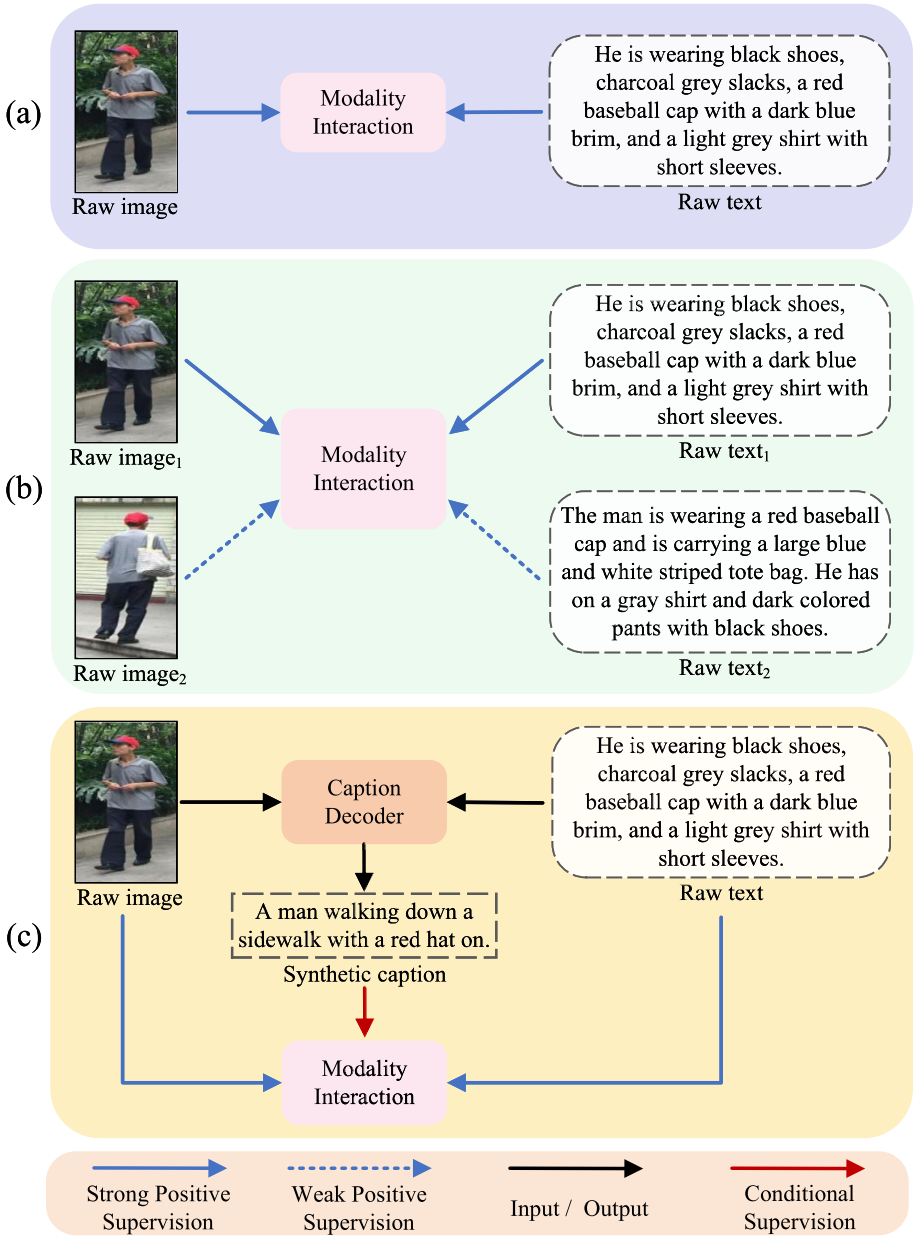


Fig. 1 Illustration of different kinds of modality interaction between multimodal data for TBPS

where a single image-text pair of the same identity is directly aligned. These methods focus on learning semantic correlations between image regions and corresponding textual phrases. However, their reliance on a strong alignment assumption makes them vulnerable to the biases and incompleteness often found in manually annotated data, leading to suboptimal retrieval performance. Therefore, the model might learn incorrect associations between image and text, leading to an inferior retrieval performance.

To this end, some recent methods [5, 6] consider many-to-many embedding framework (as illustrated in Fig. 1b), which matches images and text descriptions from multiple viewpoints. These methods extend one-to-one embedding framework by incorporating additional texts and images of the same identity through weak supervision, thereby enriching the modality interaction process. This strategy introduces complementary information to enhance modality alignment. However, the integration of additional information from the same identity inevitably leads to intra-class variation, which arises from perspectives, occlusion and so on.

In this paper, we explore the potential of synthetic captions for enhancing multi-modal representation learning in TBPS. While synthetic captions have been utilized in Image-Text Retrieval (ITR) [7], applying them directly to TBPS presents unique challenges. Unlike ITR, which typically involves brief textual descriptions, TBPS demands detailed and identity-specific captions. As shown in Fig. 2, the synthetic captions C generated by BLIP [8] often lack precision and quality,

sometimes introducing misleading details. Despite these limitations, we demonstrate that synthetic captions can provide valuable complementary information when carefully integrated into the retrieval framework.

For instance, in Fig. 2c, the caption inaccurately suggests the presence of a 'parked motorcycle' which is not depicted in the image. While synthetic captions often exhibit such issues, they still provide additional details that are valuable for TBPS, as shown in Fig. 2a and b. Therefore, it is possible to harness the useful details to enhance the accuracy and robustness of TBPS.

To this end, we propose a new embedding strategy to address the two key challenges previously identified (as shown in Fig. 1c). This strategy, which leverages synthetic captions generated from a multimodal caption decoder as conditional supervision, not only enables the model to learn more comprehensive representations but also helps to mitigate the lack of limited training data.

Our primary goal is to generate informative synthetic captions. To achieve this, we first propose an innovative adaptive masking strategy, named the Granularity Awareness Sensor (GAS), which is designed to facilitate the model's perception of fine-grained features. Overall, GAS selects fine-grained features based on the self-attention scores from the last layer of the transformer block. The process unfolds in two stages: initially, the class token's similarity with local tokens is computed to pick out the Top-K representative features. Nevertheless, these features might be disproportionately affected by attributes such as color [9] and size, which could lead the model to neglect other discriminative features. To address this, the second stage involves using the first stage's outcomes as queries to recalculate similarities with other local tokens, and choosing the token ranks first as the discriminative feature. It is worth noting that the entire mask generation process introduces no additional parameters, yet enabling the model to generate captions centered around the details about the individual.

Different from the role of synthetic caption in ITR, where aims to address low matching quality between images and texts on the internet. We shift the basis for weight allocation from captions to the similarity between image and text to learn more informative feature representations. In order to further achieve cross-modal alignment between image, text, and caption, we introduce a Conditional Contrastive Learning (CCL) which includes contrastive losses for image-text, text-caption, and image-caption pairs. (a) The image-text contrastive loss primarily enable unimodal encoders to understand the semantic meanings across different modalities. (b) The text-caption contrastive loss plays a crucial role in aligning captions with corresponding text. This alignment facilitates the generation of captions that are both accurate and enriched with a wider range of descriptions. (c) Considering the fact that captions can sometimes contain useful information not explicitly mentioned in the raw text, exclusively relying on text-caption contrastive learning might lead the model to misinterpret these useful captions as noise. Therefore, the implementation of image-caption contrastive loss is essential, which not only supplements the information that might be missing in the text but also cultivates a more comprehensive representation of the captions. (d) Finally, to tackle the challenge of potentially inaccurate captions, we propose a conditional strategy that dynamically adjusts weights within the contrastive learning framework. This adjustment is based on the similarity scores of hard negative samples, which are semantically similar yet factually unrelated. These scores serve as a practical lower bound for assessing the reliability of image-caption and text-caption pairs, aiding in mitigating the impact of inaccurate captions on retrieval performance. In summary, we highlight the contributions of this paper as follows:

- We introduce a simple but effective framework CASC for TBPS, which achieves cross-modal alignment with synthetic caption. This framework employs synthetic caption as conditional supervision and dynamically adjusts the weights of image-caption and text-caption contrastive loss based on the hard negative similarities.
- We propose an innovative adaptive masking strategy to equip the model with a refined perception for informative and discriminative features.
- We conduct extensive experiments and prove that CASC achieves state-of-the-art performance on three public benchmark datasets.

2 Related work

2.1 Vision-language pre-training

Vision-Language Pre-training Model (VLPM) has recently achieved tremendous success in TBPS task, largely attributed to the rich prior knowledge acquired from large-scale image-text pairs. Among these notable examples of VLPM, the advent of CLIP (Contrastive Language-Image Pretraining) [10] simplified the training mode of multimodal tasks, which has truly accelerated the development of TBPS. CLIP maximized the similarity between relevant image-text pairs to enhance the model's ability of learning cross-modal correspondences. ALBEF [11] integrated and adjusted image and text representations before cross-modal attention, facilitating the joint learning of visual and textual features. BLIP [8] designed a multimodal mixture of encoder-decoder architecture, which enhanced the quality of training data by generating synthetic captions and filtering out those with excessive noise. BLIP-2 [12] adopted a more concise pre-training strategy, employing existing unimodal visual and

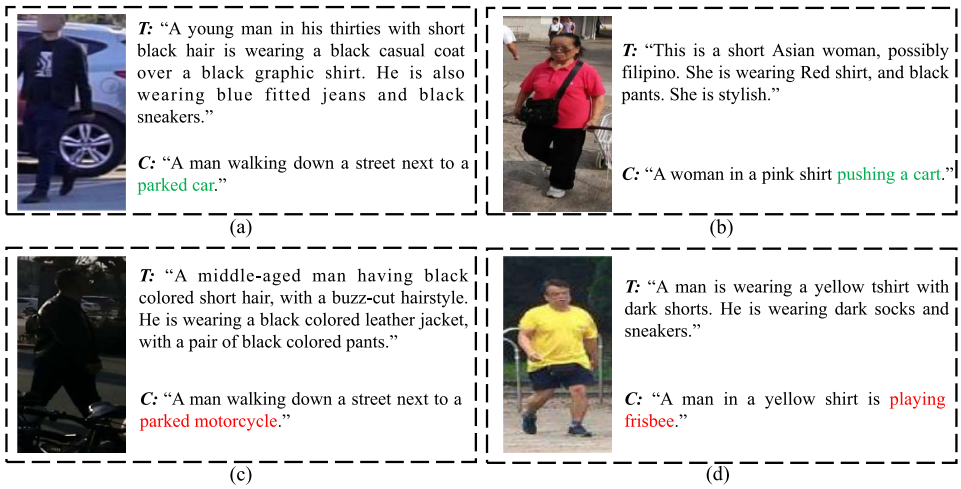


Fig. 2 Visualization of image caption generated by BLIP. T represents the raw text and C denotes the synthetic caption generated by BLIP. Green descriptions indicate useful details not mentioned in the raw text, while red descriptions descriptions represent incorrect information

textual models to minimize computational expenses and evade issues of catastrophic forgetting problems. ALIP [7] introduced an Adaptive Contrastive Loss, which dynamically adjusted the weights among image, original text, and generated synthetic caption, effectively reducing the impact of noisy image-text data on pre-training. With the continuous development in the field of VLPM, more efforts are being devoted to the design of image-grounded text decoder, aimed at providing more accurate caption descriptions for images. This progress has made the application of synthetic captions in TBPS tasks a practical possibility.

2.2 Text-based person search

Since Li et al. [13] first introduced TBPS and contributed the CUHK-PEDES dataset, this field has witnessed rapid growth in recent years. How to effectively implement modality interaction to bridge the modality gap has become a key challenge in this area. Existing methodologies for modality interaction primarily fall into two categories: one-to-one embedding and many-to-many embedding. The one-to-one embedding approach focuses on creating direct correspondences between individual text and images. For instance, ViTAA [2] adopted an auxiliary attribute segmentation layer to finely separate an identity into different attributes, aligning visual and textual attributes through contrastive learning. PMA [14] utilized human pose information to improve the alignment of multi-granularity semantic relevances between different modalities. On the other hand, many-to-many embedding strategies establish correspondences between images and texts of the same identity captured from different perspectives, aiming to capture rich image and text features from multiple viewpoints. SSAN [5] proposed Compound Ranking Loss (CR Loss), which uses textual descriptions of the same identity as weak supervision to optimize ranking performance. In addition, to alleviate intra-class variations, SSAN introduced an adaptive adjustment strategy to optimize the margin in CR Loss. Bai et al. [15] proposed a Relation and Sensitivity aware representation learning method to effectively distinguish between strong and weak positive pairs. More recently, some researchers have made new contributions in the field of TBPS by leveraging generative models. APTM [16] utilized prevalent diffusion models and image captioning models to construct a synthetic image-text dataset MALS. Similarly, Bai et al. [17] introduced a two-stage Generation-Then-Retrieval framework without the need for parallel image-text data. While these methods have made significant contributions, they rely heavily on the quality of generated captions or image-text correspondences, which can limit their effectiveness in noisy or incomplete datasets. Unlike these works, our approach, CASC, generates captions and incorporates them for cross-modal alignment under an end-to-end framework. We introduce two key innovations to address the limitations of prior approaches:

- GAS (Generation Alignment Strategy): Instead of relying on pre-generated captions or separate stages for generation and retrieval, GAS generates captions dynamically during the alignment process, ensuring better quality and relevance in real-time alignment.
- CCL (Conditional Contrastive Learning): This approach helps refine the quality of captions by adjusting contrastive loss based on similarity scores, effectively addressing the impact of low-quality captions and improving overall retrieval accuracy.

While previous methods like APTM and Bai et al. have made progress using generative models or compound ranking strategies, CASC offers a more integrated and robust solution by improving cross-modal alignment in a unified end-to-end framework.

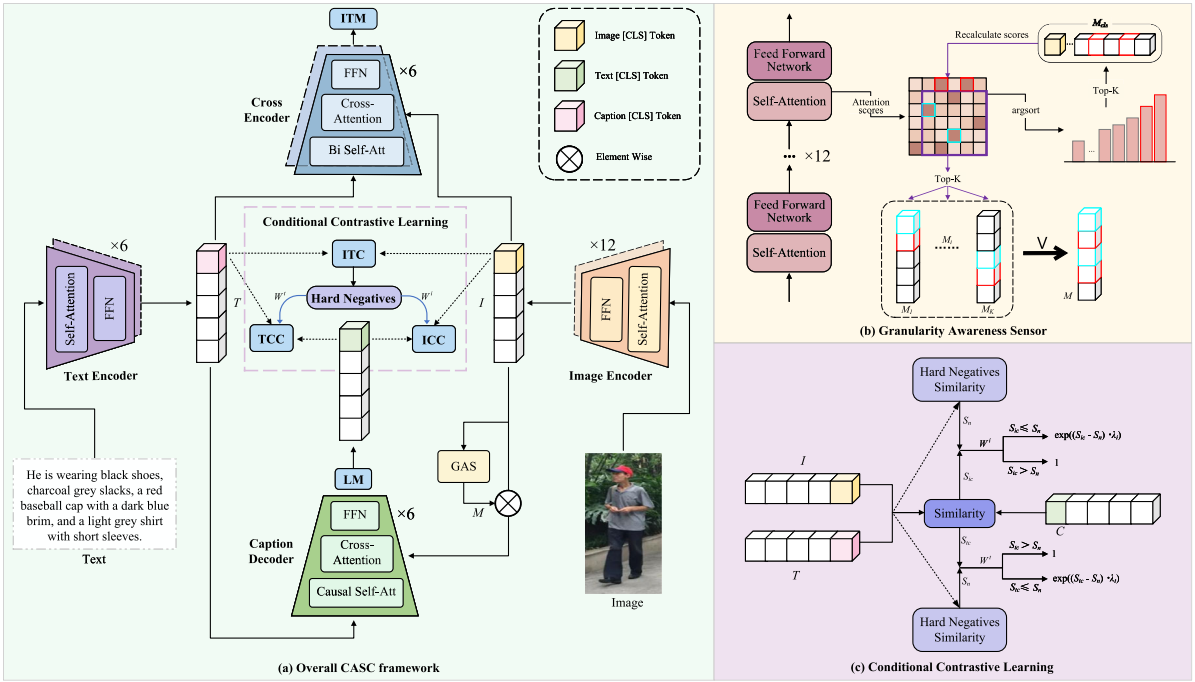


Fig. 3 The overall framework of CASC. **a** Structures of Image Encoder, Text Encoder, Cross Encoder, Caption Decoder, Granularity Awareness Sensor. **b** Conditional Contrastive Learning module. **c** Based on the discriminative and informative tokens selected from Granularity Awareness Sensor, Caption Decoder is able to generate more detailed representations. CASC employs Conditional Contrastive Learning, which comprises of image-text, image-caption and text-caption contrastive loss for cross-modal alignment. This process dynamically assigns weights by comparing hard negative similarities from image-text contrastive learning. During training, the two unimodal encoders learn more comprehensive feature representations, which facilities more accurate matching in inference

3 Methodology

3.1 Model architecture

As shown in Fig. 3, we maintain image-text retrieval architecture of BLIP [8] as the backbone, which consists of an image encoder, a text encoder, and a cross encoder. In addition, we utilize BLIP decoder as the caption decoder. Inspired by MoCo [18], we learn momentum uni-modal encoders (an exponential-moving-average version) for a consistent representation of either modality.

Image Encoder is designed as a Vision Transformer [19] comprising 12-layers of transformer blocks. Given an input image I, the encoder first decomposes it into N discrete patches. These patches are then linearly embedded, along with a learnable class token, which is denoted as $\{v_{cls}, v_1, \cdots, v_N\}$.

Text Encoder employs BERT [20] consisting of 6-layers of transformer blocks. Similar to the approach with the image encoder, we acquire a sequence of textual representations $\{t_{cls}, t_1, \cdots, t_N\}$ by feeding text T into the text encoder.

Cross Encoder is composed of an extra cross-attention layer between the self-attention layer and the feed-forward network in each transformer block for prediction.

Caption Decoder shares a similar architecture with Cross Encoder, while it uses the fused representations for generation.

3.2 Generaion

Image caption task typically uses an image-grounded text decoder constrained by Language Modeling Loss (LM) to transform visual information into coherent captions. Given an image-text pair $(I, T) \sim P$, we employ a Transformer architecture to generate captions, with T serving as query (Q) and I serving as both key (K) and value (V). This process can be expressed as follows:

$$C = Transformer(MHCA(LN(T, I, I)), \quad (1)$$

where C denotes the synthetic caption fused with image and text, $LN(\cdot)$ refers to Layer Normalization and $MHCA(\cdot)$ stands for multi-head cross attention, which can be formu-

lated as:

$$MHCA(Q, K, V) = softmax\left(\frac{QK^{\top}}{\sqrt{d}}\right)V \tag{2}$$

where d is the embedding dimension of text tokens.

The Language Modeling Loss employs a cross-entropy loss to maximize the likelihood of the raw text, which can be defined by:

$$\mathcal{L}_{lm} = \mathbb{E}_{(I,T)\sim\mathcal{P}}\mathcal{H}(y^{lm}, \psi(T \mid C)) \tag{3}$$

where $\mathcal{H}(\cdot)$ represents a cross-entropy function, y^{lm} is the one-hot vector representing true target raw text. $\psi(\cdot)$ is the probability distribution predicting the likelihood of the next word in the target T, given the caption C.

However, most image captioning models are trained on generic datasets that encompass a broad range of objects and scenes rather than specific person. TBPS requires descriptions detailed enough to distinguish individual identities. Therefore, directly applying synthetic caption for modality interaction is often inappropriate.

Granularity awareness sensor

Our primary goal is to generate more specific and detailed captions, a crucial step is to suppress noise elements to enhance the perception of fine-grained granularity with masking strategy. Traditional methods such as U-Net [21] tends to substantially increase params. In response, we introduce the Granularity Awareness Sensor (GAS), an adaptive way to generate mask without the additional parameters. Inspired by CFine [22], class token reflects the correlation with local patches, and by calculating attention scores between class token and local patches, the model can generate a mask according to the most informative tokens. The Granularity Awareness Sensor (GAS) leverages the attention scores from the last transformer block to rank and select tokens. Since this process relies entirely on pre-computed attention maps and sorting operations, it does not involve any additional trainable parameters. Formally, we initialize an image mask $M_{cls} = \{m_1, \ldots, m_N\} \in \mathbb{R}^N$ that is entirely filled with zeros, and employ $\{\boldsymbol{q}, \boldsymbol{k}\}$ from the self-attention of the last Transformer block to implement the above steps:

$$A_{cls} = \frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}} \boldsymbol{q}_{cls}(h) \cdot \boldsymbol{k}^{\top}(h) \tag{4}$$

$$m_n = \begin{cases} 1, & \text{if } n \in \text{argsort}(-A_{cls})[:K] \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

where $\mathcal{H}$ denotes the attention heads, h is an individual attention head, $\boldsymbol{q}_{\text{cls}}$ represents the cls token used as the query, $\boldsymbol{k}$ stands for other local tokens as the key, and argsort$(-(\cdot))$ refers to arranging in descending order.

However, directly selecting the Top-K tokens with the strongest responses to form a mask presents two issues: on one hand, determining an appropriate K value is inherently difficult. On the other hand, selecting top-K tokens may only highlight the most prominent features (such as clothes), while ignoring some sub-strong tokens that can also enhance individual discriminability. It is often observed that regions around the most prominent patches have higher similarities, which frequently contain informative features (such as unique jewelry). With this insight, we treat the top-K tokens as queries and the local tokens as keys to recompute similarities. From this computation, we select the Top-1 token as sub-strong tokens, which can be expressed as follows:

$$A_i = \frac{1}{|\mathcal{H}|}\sum_{h\in\mathcal{H}} \boldsymbol{q}_i(h) \cdot \boldsymbol{k}^{\top}(h), i \in K \tag{6}$$

For each strong token, its mask $M^i = \{m_1^i, \ldots, m_N^i\} \in \mathbb{R}^N, i \in K$ is acquired by:

$$m_n^i = \begin{cases} 1, & \text{if } n = \text{argsort}(-A_i)[0] \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

The final mask M is obtained by performing a logical OR operation between M_{cls} and the K individual masks M_i:

$$M = M_{cls} \vee \bigvee_{i=1}^{K} M_i \tag{8}$$

where $\vee$ represents logical OR operation.

Granularity Awareness Sensor selects tokens that represent both global and discriminative information by generating adaptive masks, enabling the model to synthesis specific and detailed caption C^m corresponding to these tokens:

$$C^m = Transformer(MHCA(LN(T, I^m, I^m)), \tag{9}$$

where $I^m = I \odot M$, $\odot$ means Hadamard product. The Language Modeling Loss of our framework can be expressed as:

$$\mathcal{L}_{lm} = \mathbb{E}_{(I,T)\sim\mathcal{P}}\mathcal{H}(y^{lm}, \psi(T \mid C^m)) \tag{10}$$

3.3 Conditional contrastive learning

Image-Text Contrastive Learning focuses on differentiating between positive and negative pairs to help two uni-modal encoders learning better representations of semantic correlation. Following MoCo [18], we maintain momentum-based uni-modal encoders to store a larger number of negative samples, and employ the InfoNCE [23] loss to optimize the alignment between different modalities. Specifically, given

an image-text pair $(I, T) \sim P$ and two queues with a capacity of K, the image-to-text contrastive loss can be formulated by:

$$\mathcal{L}_{i2t}\left(I, T^{+}, T^{-}\right) = -\mathbb{E}_{p(I,T)}\left[\log \frac{\exp\left(S\left(I, T^{+}\right)/\tau\right)}{\sum_{k\in K}\exp\left(S\left(I, T_{k}^{-}\right)/\tau\right)}\right] \tag{11}$$

where $T^{-} = \left\{T_{1}^{-}, \ldots, T_{K}^{-}\right\}$ refers to the negative text samples stored in the queue, T^{+} is the positive text sample that correctly matches with I, τ is a temperature hyper-parameter. We use two projection heads h_v and $\hat{h}_t$ to map visual representation v_{cls} and momentum textual representation $\hat{h}_{cls}$ to a lower-dimensional (256) for similarity calculation: $S(I, T) = h_v\left(v_{\text{cls}}\right)^{\mathrm{T}} \hat{h}_t\left(\hat{t}_{\text{cls}}\right)$. Likewise, text-to-image contrastive loss is as follow:

$$\mathcal{L}_{t2i}\left(T, I^{+}, I^{-}\right) = -\mathbb{E}_{p(I,T)}\left[\log \frac{\exp\left(S\left(T, I^{+}\right)/\tau\right)}{\sum_{k\in K}\exp\left(S\left(T, I_{k}^{-}\right)/\tau\right)}\right] \tag{12}$$

We incorporate all the losses to form Contrastive Loss as follows:

$$\mathcal{L}_{itc} = \frac{1}{2}\left[\mathcal{L}_{i2t}\left(I, T^{+}, T^{-}\right) + \mathcal{L}_{t2i}\left(T, I^{+}, I^{-}\right)\right] \tag{13}$$

To further mitigate the negative impact of inaccurate or noisy captions while ensuring the semantically alignment between synthetic captions and the image-text paris, we introduce a conditional contrastive loss within our contrastive learning framework. In practice, We adopt a strategy for hard negative sampling similar to ALBEF [11]. Considering in a mini-batch, we sample its negative image according to the similarity of S^{t2i} and pick the highest one as the hard negative scores S_n^{t2i}. The negative text S_n^{i2t} is sampled by the similar manner. We choose the smaller of the two scores as the standard:

$$S_n = min(S_n^{t2i}, S_n^{i2t}) \tag{14}$$

To accomplish the adjustment, we design two dynamic sample weights W^i and $W^t \in (0, 1]$ given by the following equation:

$$W^{i} = \begin{cases} \exp\left(\left(S_{ic} - S_n\right) * \lambda_i\right), & S_{ic} \le S_n \\ 1, & S_{ic} > S_n \end{cases}, \tag{15}$$

$$W^{t} = \begin{cases} \exp\left(\left(S_{tc} - S_n\right) * \lambda_t\right), & S_{tc} \le S_n \\ 1, & S_{tc} > S_n \end{cases}, \tag{16}$$

where λ_i and λ_t are hyper-parameters, S_{tc} is the similarity between raw text and synthetic caption and S_{ic} is the similarity between image and synthetic caption.

According to the weights assigned to the contrastive learning, given a triplet (I, T, C), the image-caption and text-caption can be defined as follows:

$$\mathcal{L}_{icc} = \frac{1}{2} W^{i}\left[\mathcal{L}_{i2c}\left(I, C^{+}, C^{-}\right) + \mathcal{L}_{c2i}\left(C, I^{+}, I^{-}\right)\right] \tag{17}$$

$$\mathcal{L}_{tcc} = \frac{1}{2} W^{t}\left[\mathcal{L}_{t2c}\left(T, C^{+}, C^{-}\right) + \mathcal{L}_{c2t}\left(C, T^{+}, T^{-}\right)\right] \tag{18}$$

Finally, the Conditional Contrastive Loss is formulated as:

$$\mathcal{L}_{ccl} = \mathcal{L}_{itc} + \mathcal{L}_{icc} + \mathcal{L}_{tcc} \tag{19}$$

3.4 Optimization & inference

Optimization. In addition to Conditional Contrastive Learning (CCL) and Language Modeling (LM), CASC is also supervised by Image-Text Matching (ITM) during training.

Image-TextMatching focuses on predicting whether an image-text pair is matched. Following [11], We adopt the hard negative sampling strategy to provide more informative training examples for our training process. The ITM loss is calculated using both positive image-text pairs and these specially selected negative pairs, processed through a Cross Encoder and prediction heads, which can be defined as:

$$\mathcal{L}_{itm} = \mathbb{E}_{p(I,T)}\mathcal{H}\left(y^{itm}, \phi^{itm}(I, T)\right) \tag{20}$$

where $\mathcal{H}(\cdot)$ represents a cross-entropy function, y^{itm} is a 2-dimensional one-hot vector denoting the ground-truth label and $\phi^{itm}(I, T) = Transformer(MHCA(LN(T, I, I))$. Given the above optimization objectives, the full architecture loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{itm} + \mathcal{L}_{ccl} + \mathcal{L}_{lm} \tag{21}$$

Inference. Since datas are not in paired relationships during inference and the model has already learned additional supervision information, we only use two unimodal encoders and a cross encoder for inference. For each image, the model computes similarity scores with all text representations, typically using the joint representations from the cross encoder. The higher the similarity score, the more relevant the text is considered to be for the given image.

4 Experiments

We conduct the experiments on three benchmark datasets, with BLIP as backbone model.

4.1 Datasets and evaluation metrics

CUHK-PEDES is a commonly-used dataset dataset for text-based person search, comprising 40,206 images and 80,440 textual descriptions across 13,003 identities. It splits into 34,054 images and 68,126 texts for training, 3078 images and 6158 texts for validation, and 3074 images with 6156 texts for testing. Each descriptive text averages around 23 words, focusing on the distinctive appearance of each individual depicted in the dataset.

ICFG-PEDES is sourced from MSMT17, which contains 54,522 images paired with textual descriptions for 4102 unique identities. The dataset is partitioned into a training set with 34,674 images of 3102 identities, and a test set featuring 19,848 images of 1000 identities. Each description in this dataset averages around 37 words.

RSTPReid is the recent benchmark focusing on real-world scenarios, constructed from the MSMT17 dataset. It includes 20,505 images and 41,010 textual descriptions representing 4101 individuals. There are 3, 701/200/200 identities utilized for training/validation/testing, respectively.

Rank-K is a metric that evaluates the accuracy of retrieval results. Specifically, it checks whether the correct match (the image matching the description, or vice versa) appears in the top K results of a search query.

Mean Average Precision (mAP) provides a more comprehensive assessment by considering both precision and recall across all queries. It calculates the average precision at different recall levels for each query and then averages these values across all queries.

4.2 Implementation details

All of our experiments are performed on 4 NVIDIA A100 GPUs with PyTorch framework. We optimize the CASC model for 30 epochs with a batch size of 64. AdamW optimizer is adopted with a weight decay of 0.05. We utilize the BLIP Image-Text Retrieval (Flicker30K) backbone, while the decoder is initialized with pre-trained weights from Image Captioning (COCO). For image processing, we resize all inputs to 384 $\times$ 384 pixels and apply random horizontal flips as part of our data augmentation strategy. The model employs a queue size K of 57,600 and a contrastive learning temperature τ of 0.07. To maintain consistency in representation, the momentum model's coefficient is fixed at 0.995. Furthermore, in our approach, we define a threshold of the Top-40 tokens (K set to 45) for selecting discriminative features the Granularity Awareness Sensor. Following ALIP, λ_i and λ_t are set to 2.

Table 1 Performance comparison on CUHK-PEDES dataset

	Method	R@1	R@5	R@10	mAP
w/o VLP	ViTAA [2]	55.97	75.84	83.52	–
	DSSL [24]	59.98	80.41	87.56	–
	SSAN [5]	61.37	80.15	86.73	–
	LapsCore [25]	63.40	–	87.80	–
	SAF [26]	64.13	82.62	88.40	58.61
	LBUL [27]	64.04	82.66	87.22	–
	CAIBC [9]	64.43	82.87	88.37	–
	LGUR [28]	65.25	83.12	89.00	–
	AXM-Net [29]	64.44	80.52	86.77	58.73
	BEAT [6]	64.23	82.91	88.65	–
	LCR^{2}S [30]	67.36	84.19	89.62	59.24
w/ VLP	IVT [31]	65.59	83.11	89.21	58.99
	CFine [31]	69.57	85.93	91.15	–
	TCB [32]	74.45	90.07	**94.66**	64.12
	APTM [16]	76.53	90.04	94.15	66.91
	IRRA [33]	73.38	89.93	93.71	66.13
	RaSa [15]	76.51	90.29	94.25	**69.38**
	TBPS-CLIP [34]	73.54	88.19	92.35	65.38
	BiLMa [35]	74.03	89.59	93.62	66.57
	RDE [36]	75.94	90.14	94.12	67.56
	PP [37]	74.89	89.90	94.17	67.12
	AUL [38]	77.23	90.43	94.41	–
	BLIP (Backbone)	65.61	82.84	88.65	58.02
	CASC	**77.71**	**90.57**	94.22	69.16

Bold values indicate that the corresponding value is the highest among the compared data points, thereby highlighting the most significant result

4.3 Comparison with state-of-the-art methods

As illustrated in Tables 1, 2 and 3 respectively, our CASC outperforms the state-of-the art methods on CUHK-PEDES [13], ICFG-PEDES [5] and RSTPReid [24] in terms of R@1. Specifcally, we first compare our method with the existing best-performing model RaSa [15], which outperforms +1.58%, +4.57% and +3.90% on three datasets in R@1, respectively. RaSa still belongs to many-to-many embedding works, detecting word replacements to enhance the model's ability to perceive subtle differences in textual descriptions. This strategy is also dedicated to mitigate the impacts of the noise interference from weak positive pairs. Our CASC considers TBPS as a new embedding perspective, augmenting the diversity of textual descriptions with synthetic captions to facilitate learning more generalized representations. We fur-

Table 2 Performance comparison on ICFG-PEDES dataset

	Method	R@1	R@5	R@10	mAP
w/o VLP	Dual Path [39]	38.99	59.44	68.41	–
	CMPM/C [40]	43.51	65.44	74.26	
	ViTAA [2]	50.98	68.79	75.78	
	SSAN [5]	54.23	72.63	79.53	–
	SAF [26]	54.86	72.13	79.13	32.76
	TIPCB [41]	54.96	74.72	81.89	–
	SRCF [42]	57.18	75.01	81.49	–
	LGUR [28]	59.02	75.32	81.56	–
	BEAT [6]	58.16	75.91	82.04	–
	LCR^{2}S [30]	57.93	76.08	82.40	38.21
w/ VLP	IVT [31]	56.04	73.60	80.22	–
	CFine [31]	60.83	76.55	82.42	–
	TCB [32]	61.60	76.33	81.90	**44.31**
	APTM [16]	68.51	82.99	**87.56**	41.22
	IRRA [33]	63.46	80.25	85.82	38.06
	RaSa [15]	65.28	80.40	85.12	41.29
	TBPS-CLIP [34]	65.05	80.34	85.47	39.83
	BiLMa [35]	63.83	80.15	85.74	38.26
	RDE [36]	67.68	82.47	87.36	40.06
	PP [37]	65.12	81.57	86.97	42.93
	BLIP (Backbone)	37.09	55.19	63.65	21.39
	CASC	**69.85**	**84.03**	86.79	43.58

Bold values indicate that the corresponding value is the highest among the compared data points, thereby highlighting the most significant result

Table 3 Performance comparison on RSTPReid dataset

	Method	R@1	R@5	R@10	mAP
w/o VLP	AMEN [43]	38.45	62.40	73.80	–
	LBUL [27]	45.55	68.20	77.85	–
	DSSL [24]	39.05	62.60	73.95	
	SSAN [5]	43.50	67.80	77.15	–
	SAF [26]	44.05	67.30	76.25	36.81
	CAIBC [9]	47.35	69.55	79.00	–
	BEAT [6]	46.90	70.90	79.35	–
	LCR^{2}S [30]	54.95	76.65	84.70	40.92
w/ VLP	IVT [31]	49.70	70.00	78.80	–
	CFine [31]	50.55	72.50	81.60	–
	TCB [32]	65.80	82.85	88.20	52.70
	APTM [16]	68.51	82.99	87.56	41.22
	IRRA [33]	60.20	81.30	88.20	47.17
	RaSa [15]	66.90	86.50	91.35	52.31
	TBPS-CLIP [34]	61.95	83.55	88.75	48.26
	BiLMa [35]	61.20	81.50	88.80	48.51
	RDE [36]	65.35	83.95	89.90	50.88
	PP [37]	61.87	83.63	89.70	47.82
	AUL [38]	69.16	83.32	88.37	–
	BLIP (Backbone)	58.25	77.85	85.65	44.08
	CASC	70.80	87.35	93.25	55.39

Table 4 Comparison of different metrics across models

	C ⇒ I	I ⇒ C	C ⇒ R	R ⇒ C	I ⇒ R	R ⇒ I
TCB	33.96	39.05	49.40	34.93	**51.75**	39.77
CASC	**45.76**	**45.33**	**51.08**	**38.42**	50.26	**42.84**

Bold values indicate that the corresponding value is the highest among the compared data points, thereby highlighting the most significant result

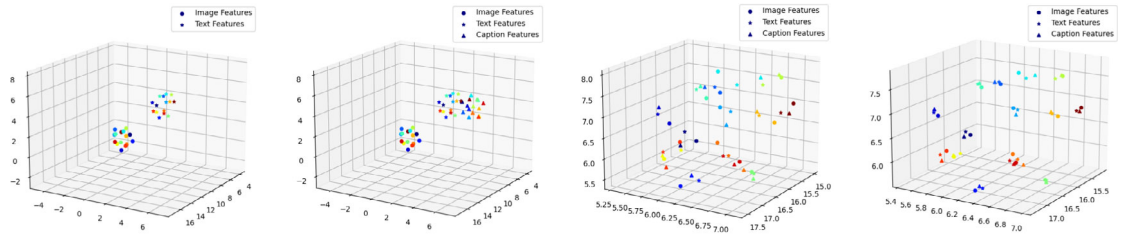
ther compare our proposed method with APTM [16], which is formally similar to our work. Compared to APTM, which captures more diverse data representations through Masked Attribute Modeling and proposed image-attribute prompts contrastive learning to facilitate the representation learning, we directly use Language Modeling to generate coherent captions. Moreover, they didn't consider the correctness of the generated prompts, assigning the same weight to all in the contrastive learning process, resulting in suboptimal performance compared to CASC (+1.18%, +2.35% and +2.29%).

4.4 Comparison with cross domain performance

To address performance beyond tested datasets, here, we conduct cross-domain experiments among three datasets. The Rank-1 results compared to TCB [32] are shown in Fig. 4. The comparison can clearly demonstrate that, due to our model learning additional and effective textual information, there is a significant improvement in generalization performance Table 4.

4.5 Ablation study

Effectiveness of Optimization Objectives. CASC consists of five Optimization Objectives. All ablation experiments are shown in Table 5. First, we maintain the original ITM and ITC loss to fine-tune the existing BLIP model for our specific downstream task. ITC and ITM have effectively learned feature alignment and interaction, resulting in a notable increase from 65.61% to 73.26% in R@1 in the comparison of No.0 vs No.1. Subsequently, in No.2 we added the caption decoder, combined with LM loss to the baseline, which were not directly involved in the retrieval process, resulting in a minor performance reduction from No.1. However, with the incorporation of text-caption, image-caption, and their combined optimization in No.3, No.4, and No.5. We noticed an approximate 1–2% improvement in R@1 over the results of No.2 This advancement highlights the significant role of addi-



(a) Baseline (b) Baseline w/ CD (c) CASC w/o CCL (d) CASC

Fig. 4 3D UMAP visualizations of feature embeddings for different model configurations on ICFG-PEDES: **a** Baseline model, **b** Baseline model with Caption Decoder (CD), **c** CASC without Conditional Contrastive Learning (CCL), and **d** Our proposed CASC

tional caption supervision in developing a more detailed and comprehensive feature representation, which significantly contributes to retrieval effectiveness.

Effectiveness of Granularity Awareness Sensor. GAS leverages off-the-shelf Transformer model features and self-attention mechanisms to help the model adaptively understand and generate more accurate and detailed descriptions without introducing additional parameters. In contrast to directly selecting the Top-2K tokens, GAS incorporates both Top-K tokens and secondary strong tokens, enabling the model to capture a more diverse and detailed set of image features. This approach allows GAS to extend its focus beyond the most prominent elements, ensuring finer granularity in feature selection. The experimental results comparing No.5 vs. No.6 clearly demonstrate the efficacy of GAS. When GAS is added on top of the Caption Decoder and contrastive learning, a 1.08% improvement in the R@1 metric is observed. To further substantiate this improvement, we present comprehensive visualizations that highlight the differences between various masking strategies. Figure 6 illustrates the comparison between Top-2K, Top-K, and GAS masking strategies.

- **Top-2K:** The Top-2K strategy focuses primarily on the most prominent image features, often leading to an overemphasis on edges and background details, which can introduce noise into the feature selection process.
- **Top-K:** The Top-K strategy improves upon this by selecting a larger set of features but still tends to prioritize more general features, potentially overlooking finer details in the image.
- **GAS:** In contrast, GAS not only emphasizes the central elements of the image, such as the individual's face

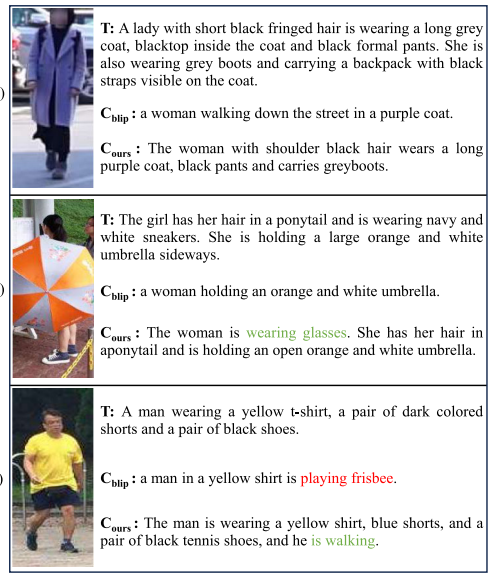


Fig. 5 Comparisons of captions generated from BLIP and CASC

Table 5 Ablation study on different components of CASC on CUHK-PEDES dataset

No	Method	$\mathcal{L}_{itc}$	$\mathcal{L}_{itm}$	$\mathcal{L}_{lm}$	$\mathcal{L}_{tcc}$	$\mathcal{L}_{icc}$	R@1	R@5	R@10
1	BLIP						65.61	82.84	88.65
2	Baseline	✓	✓				73.26	88.12	92.22
3		✓	✓	✓			72.35	88.19	91.89
4		✓	✓	✓	✓		74.09	89.46	93.53
5		✓	✓	✓		✓	73.51	89.37	93.10
6	+CD	✓	✓	✓	✓	✓	75.09	89.96	**94.53**
7	+CD+GAS	✓	✓	✓	✓	✓	76.17	90.12	93.88
8	+CD+GAS+CCL	✓	✓	✓	✓	✓	**77.71**	**90.57**	94.22

Bold values indicate that the corresponding value is the highest among the compared data points, thereby highlighting the most significant result

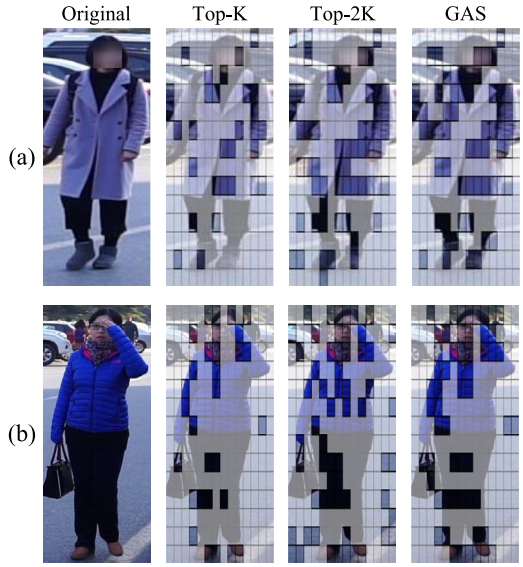


Fig. 6 Comparisons of different masking strategy, where the darkened areas indicate the masked tokens, here $K = 30$

or body, but also effectively captures small yet crucial details, such as shoes in (a) and a handbag in (b). This enhanced focus on detailed features is particularly important for tasks that require high precision and an understanding of subtle visual cues.

Effectiveness of Conditional Contrastive Learning. The comparison between Experiment No.7 and No.8 directly validates the effectiveness of our proposed Conditional Contrastive Learning, as evidenced by a 1.54% increase in the R@1 metric. We visualized the changes in hard negative similarities as well as image-caption and text-caption similarities during training process. As demonstrated in Fig. 7.

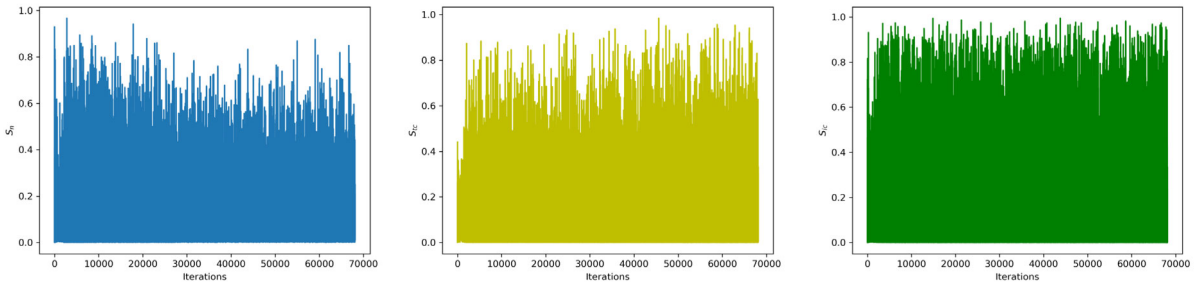
Ablation Study of selection of Top-K tokens. The selection of Top-K tokens has a certain impact on the experimental results, as shown in Table 6. Too few tokens cannot encapsulate individual detailed features, while too many tokens inevitably introduce background information. We therefore chose a balanced number of tokens to capture the essential details while minimizing the inclusion of irrelevant background data.

4.6 Visualization

UMAP Visualization. We randomly sampled 15 image-text pairs and used UMAP [44] for visualization, as shown in Fig. 4. The UMAP plot demonstrates the distribution and clustering behavior of image and text features in our model's embedding space.

- **Caption Decoder Incorporation:** When the caption decoder is incorporated into the baseline, the features generated through cross-attention mechanisms between the image and text modalities are semantically similar to the text features but not identical. This is visually represented in Fig. 4b, where caption features are positioned near, but not exactly on top of, the text features, indicating that the model has learned a shared semantic space but still requires further optimization for fine-grained alignment.
- **Joint Contrastive Learning Optimization:** As we optimize the model using joint contrastive learning across image-text, image-caption, and text-caption pairs, features of the same identity gradually draw closer together in the feature space. This is evident in the plot as the features from the same identity form tighter clusters, with reduced overlap between different identities. The inclusion of conditional contrastive learning further enhances this clustering, allowing features of the same identity to form well-defined clusters that are distinctly separated from other identities, as shown by the clear separations in the plot.

Synthetic Caption Visualization. To demonstrate the informativeness of our synthetic captions, we decoded their



(a) S_n (b) S_{tc} (c) S_{ic}

Fig. 7 Visualization of different similarities during training

Table 6 Experiments of different values of K on CUHK-PEDES

Method	Rank-1	Rank-5	Rank-10
K = 30	69.07	86.84	92.10
K = 35	72.34	88.37	92.66
K = 40	75.05	89.67	93.89
K = 45	**77.71**	**90.57**	**94.22**
K = 50	75.90	90.04	94.17
K = 60	74.27	90.03	93.89

Bold values indicate that the corresponding value is the highest among the compared data points, thereby highlighting the most significant result

feature representations to a visualization process, the results of which are depicted in Fig. 5. Figure 5 (a) indicates that compared to BLIP, our captions are more detailed, providing additional fine-grained features. Figure 5 (b) shows that our captions offer beneficial information not present in the raw text. Figure 5c illustrates that with image guidance, our captions have corrected erroneous information and accurately reflected the actions of the individuals.

5 Conclusion

In this paper, we propose a cross-modal alignment with synthetic caption framework (CASC) for TBPS. CASC focuses on synthesizing fine-grained and informative captions to achieve modality alignment between image, text, and caption. The model mainly contains two components, named GAS and CCL. Concretely, GAS facilitates the identification of discriminative features through an adaptive masking strategy, while CCL aligns different modalities through further constraints on the synthetic captions. Extensive experiments on multiple benchmarks and visualization of experiment results demonstrate the effectiveness of CASC.

Author Contributions W.Z. designed the experiments and wrote the main manuscript. Y.L. provided the experimental device (GPU). Y.Y. was responsible for designing the figures. Z.L. and G.J. reviewed and edited the manuscript. All authors reviewed the final manuscript.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interests All authors certify that they have no affiliations with or involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. Han X, He S, Zhang L, Xiang T (2021) Text-based person search with limited data
2. Wang Z, Fang Z, Wang J, Yang Y (2020) 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. Springer
3. Niu K, Huang Y, Ouyang W, Wang L (2020) Improving description-based person re-identification by multi-granularity image-text alignments. IEEE Trans Image Process 29:5542–5556
4. Ji Z, Hu J, Liu D, Wu LY, Zhao Y (2022) Asymmetric cross-scale alignment for text-based person search. IEEE Trans Multimed 25:7699–7709
5. Ding Z, Ding C, Shao Z, Tao D (2021) Semantically self-aligned network for text-to-image part-aware person re-identification. arXiv preprint arXiv:2107.12666
6. Ma Y, Sun X, Ji J, Jiang G, Zhuang W, Ji R (2023) 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
7. Yang K, Deng J, An X, Li J, Feng Z, Guo J, Yang J, Liu T (2023) Alip: Adaptive language-image pre-training with synthetic caption. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 2922–2931
8. Li J, Li D, Xiong C, Hoi S (2022) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, pp 12888–12900. PMLR
9. Wang Z, Zhu A, Xue J, Wan X, Liu C, Wang T, Li Y (2022) Caibc: Capturing all-round information beyond color for text-based person retrieval. In: Proceedings of the 30th ACM International Conference on Multimedia, pp 5314–5322
10. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, *et al* (2021) Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp 8748–8763. PMLR
11. Li J, Selvaraju R, Gotmare A, Joty S, Xiong C, Hoi SCH (2021) Align before fuse: vision and language representation learning with momentum distillation. Adv Neural Inform Process Syst 34:9694–9705
12. Li J, Li D, Savarese S, Hoi S (2023) Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning, pp 19730–19742. PMLR
13. Li S, Xiao T, Li H, Zhou B, Yue D, Wang X (2017) Person search with natural language description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1970–1979
14. Jing Y, Si C, Wang J, Wang W, Wang L, Tan T (2020) Pose-guided multi-granularity attention network for text-based person search. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp 11189–11196
15. Bai Y, Cao M, Gao D, Cao Z, Chen C, Fan Z, Nie L, Zhang M (2023) Rasa: Relation and sensitivity aware representation learning for text-based person search. arXiv preprint arXiv:2305.13653
16. Yang S, Zhou Y, Zheng Z, Wang Y, Zhu L, Wu Y (2023) Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In: Proceedings of the 31st ACM International Conference on Multimedia, pp 4492–4501
17. Bai Y, Wang J, Cao M, Chen C, Cao Z, Nie L, Zhang M (2023) Text-based person search without parallel image-text data. arXiv preprint arXiv:2305.12964

18. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9729–9738
19. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S. et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
20. Devlin J, Chang M-W, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
21. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pp 234–241. Springer
22. Yan S, Dong N, Zhang L, Tang J (2023) Clip-driven fine-grained text-image person re-identification. IEEE Trans Image Process 32:6032–6046
23. Oord Avd, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748
24. Zhu A, Wang Z, Li Y, Wan X, Jin J, Wang T, Hu F, Hua G (2021) Dssl: Deep surroundings-person separation learning for text-based person retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia, pp 209–217
25. Wu Y, Yan Z, Han X, Li G, Zou C, Cui S (2021) Lapscore: language-guided person search via color reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 1624–1633
26. Li S, Cao M, Zhang M (2022) 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
27. Wang Z, Zhu A, Xue J, Wan X, Liu C, Wang T, Li Y (2022) Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In: Proceedings of the 30th ACM International Conference on Multimedia, pp 1984–1992
28. Shao Z, Zhang X, Fang M, Lin Z, Wang J, Ding C (2022) Learning granularity-unified representations for text-to-image person re-identification. In: Proceedings of the 30th ACM International Conference on Multimedia, pp 5566–5574
29. Farooq A, Awais M, Kittler J, Khalid SS (2022) Axm-net: Implicit cross-modal feature alignment for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp 4477–4485
30. Yan S, Dong N, Liu J, Zhang L, Tang J (2023) 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
31. Shu X, Wen W, Wu H, Chen K, Song Y, Qiao R, Ren B, Wang X (2022) See finer, see more: Implicit modality alignment for text-based person retrieval. In: European Conference on Computer Vision, pp 624–641. Springer
32. Zang X, Gao W, Li G, Fang H, Ban C, He Z, Sun H (2023) A baseline investigation: Transformer-based cross-view baseline for text-based person search. In: Proceedings of the 31st ACM International Conference on Multimedia, pp 7737–7746
33. Jiang D, Ye M (2023) Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 2787–2797
34. Cao M, Bai Y, Zeng Z, Ye M, Zhang M (2023) An empirical study of clip for text-based person search. arXiv preprint arXiv:2308.10045
35. Fujii T, Tarashima S (2023) Bilma: Bidirectional local-matching for text-based person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 2786–2790
36. Qin Y, Chen Y, Peng D, Peng X, Zhou JT, Hu P (2024) Noisy-correspondence learning for text-to-image person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 27197–27206
37. Yan S, Liu J, Dong N, Zhang L, Tang J (2024) Prototypical prompting for text-to-image person re-identification. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp 2331–2340
38. Li S, He C, Xu X, Shen F, Yang Y, Shen HT (2024) Adaptive uncertainty-based learning for text-based person retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp 3172–3180
39. Zheng Z, Zheng L, Garrett M, Yang Y, Xu M, Shen Y-D (2020) Dual-path convolutional image-text embeddings with instance loss. ACM Trans Multimed Comput Commun Appl (TOMM) 16(2):1–23
40. Zhang Y, Lu H (2018) Deep cross-modal projection learning for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 686–701
41. Chen Y, Zhang G, Lu Y, Wang Z, Zheng Y (2022) Tipcb: a simple but effective part-based convolutional baseline for text-based person search. Neurocomputing 494:171–181
42. Suo W, Sun M, Niu K, Gao Y, Wang P, Zhang Y, Wu Q (2022) A simple and robust correlation filtering method for text-based person search. In: European Conference on Computer Vision, pp 726–742. Springer
43. Wang Z, Xue J, Zhu A, Li Y, Zhang M, Zhong C (2021) Amen: Adversarial multi-space embedding network for text-based person re-identification. In: Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part II 4, pp 462–473. Springer
44. McInnes L, Healy J, Melville J (2018) Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426