

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository: <https://orca.cardiff.ac.uk/id/eprint/189556/>

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Liu, Jing, Wei, Donglai, Liu, Yang, Zhang, Sipeng, Yang, Tong, Zhou, Wei, Ding, Weiping and Leung, Victor C.M. 2027. SCMM: Calibrating cross-modal representations for text-based person search. Pattern Recognition 182 , 114775. 10.1016/j.patcog.2026.114775

Publishers page: <https://doi.org/10.1016/j.patcog.2026.114775>

Please note:

Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See <http://orca.cf.ac.uk/policies.html> for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.





SCMM: Calibrating cross-modal representations for text-based person search[☆]

Jing Liu^{a,b}, Donglai Wei^{a,*}, Yang Liu^d, Sipeng Zhang^c, Tong Yang^c, Wei Zhou^e, Weiping Ding^{f,g}, Victor C.M. Leung^{h,b}

^a College of Future Information Technology, Fudan University, Shanghai, 200433, China

^b Department of ECE, The University of British Columbia, Vancouver, V6T 1Z4, Canada

^c MEGVII Technology, Beijing, 100096, China

^d College of Electr. and Info. Eng., Tongji University, Shanghai, 201804, China

^e School of Computer Science and Informatics, Cardiff University, Cardiff, CF24 4AG, UK

^f School of AI and CS, Nantong University, Nantong, 226019, China

^g Faculty of Data Science, City University of Macau, 999078, Macao Special Administrative Region of China

^h Academy of Artificial Intelligence, SMBU, Shenzhen, 518115, Guangdong, China

ARTICLE INFO

Keywords:

Text-based person search
Cross-modal learning
Multimodal perception
Pattern recognition
Metric learning
Visual-textual matching

ABSTRACT

Text-Based Person Search (TBPS) aims to retrieve target person images from a large-scale database using natural language descriptions, serving as a critical task in multimodal perception and visual pattern recognition. Bridging the semantic gap between heterogeneous modalities while capturing fine-grained correspondences remains a fundamental challenge, especially when discriminating visually similar individuals based on complex textual semantics. To address these challenges, we propose Sew Calibration and Masked Modeling (SCMM), a unified framework that calibrates cross-modal representations for effective multimodal visual-textual pattern matching. Concretely, SCMM introduces two principal components: a sew calibration loss that dynamically aligns image-text features via a quality-guided adaptive margin governed by textual information density, and a masked caption modeling loss that establishes fine-grained semantic correspondences through transformer-based masked prediction. The sew calibration mechanism imposes bidirectional constraints to compactly cluster same-identity features in a shared embedding space. Simultaneously, the masked modeling component acts as a cross-modal decoder that learns word-level representations, effectively discriminating subtle attribute differences. Importantly, our dual-encoder architecture strikes an optimal balance between representation expressiveness and computational efficiency by adopting a training-only decoder design. Extensive experiments on CUHK-PEDES, ICFG-PEDES, and RSTPReID datasets demonstrate that SCMM achieves state-of-the-art performance with Rank-1 accuracies of 73.81%, 64.25%, and 57.35%, respectively. Thorough ablation studies confirm the efficacy of each proposed mechanism in establishing robust cross-modal patterns for multimodal perception and recognition.

1. Introduction

Cross-modal data retrieval and multimodal perception have emerged as fundamental research directions in pattern recognition, driven by the increasing demand to efficiently process and understand heterogeneous data sources [1]. Person re-identification (Re-ID) exemplifies a challenging multimodal matching task that matches individuals across non-overlapping camera views by mining discriminative visual representations [2]. However, conventional Re-ID systems rely

solely on visual queries, which inherently limits their applicability when visual references are unavailable in the database [3]. Text-Based Person Search (TBPS) overcomes this limitation by employing natural language descriptions as database queries, necessitating sophisticated knowledge engineering techniques to align visual and textual modalities in a shared semantic manifold [4]. TBPS requires models to learn robust multimodal embeddings that capture both global semantic alignments and fine-grained data correspondences [5], thereby presenting unique challenges for representation learning and multimodal pattern analysis [6]. Edge-cloud collaborative intelligence further shows

[☆] This article is part of a Special issue entitled: 'PR_Multimodal Perception and Pattern Recognition' published in Pattern Recognition.

* Corresponding author.

E-mail addresses: jingliu@ece.ubc.ca (J. Liu), dlwei21@m.fudan.edu.cn (D. Wei), yang.liu@ieee.org (Y. Liu), zhangsipeng@megvii.com (S. Zhang), yangtong@megvii.com (T. Yang), zhouw26@cardiff.ac.uk (W. Zhou), dlwei21@m.fudan.edu.cn (W. Ding), vleung@ece.ubc.ca (V.C.M. Leung).

<https://doi.org/10.1016/j.patcog.2026.114775>

Received 16 May 2026; Received in revised form 18 July 2026; Accepted 18 August 2026

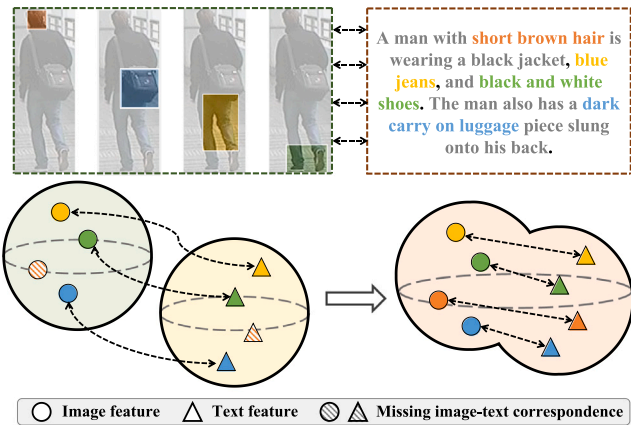


Fig. 1. Illustration of the motivation behind our method. In cross-modal tasks, a compact and well-aligned image-text feature distribution in the shared embedding space is crucial for bridging the inter-modal gap. Additionally, capturing fine-grained image-text correspondences is equally vital to distinguish between similar individuals for text-based person searching.

that multimodal representation learning is increasingly shaped by distributed deployment and model-optimization constraints [7].

The effectiveness of these retrieval systems heavily depends on bridging the semantic gap between heterogeneous modalities while preserving discriminative features essential for accurate retrieval [8]. Video anomaly detection studies similarly highlight the importance of reliable visual evidence in networked safety-critical perception systems [9]. The primary challenge in TBPS arises from the semantic discontinuity between raw visual data and structured textual concepts, requiring features to be aligned across disparate embedding distributions [10]. Large-scale clustering studies further highlight the strong value of globally organized representation spaces when retrieval systems must remain reliable at scale [11]. Recent advances in transformer-based architectures demonstrate that cross-modal attention mechanisms can facilitate rich feature interactions and semantic fusion between modalities [12]. Contemporary pattern recognition research highlights the importance of developing efficient alignment strategies that combine complementary information while ensuring computational tractability for real-time applications [13]. These advances are critical for cross-modal retrieval tasks that distinguish subtle visual differences guided by complex textual descriptions [14]. Broader machine-learning studies also motivate the need for efficient, stable, and interpretable representation learning. GPU-accelerated optimal tree learning has clearly shown how scalable optimization can effectively improve discriminative decision structures on large datasets [15]. In industrial monitoring, Cao et al. [16] made a valuable contribution to reliable soft sensing by emphasizing that accurate models should remain transparent under practical constraints.

Compared to conventional Re-ID, the main objective of TBPS is to learn fine-grained cross-modal representations between visual and textual modalities [17]. As shown in Fig. 1, cross-modal representations have two key characteristics that contribute to effectively searching for a person from images: closely aligned cross-modal representations and fine-grained information correspondence. First, closely aligned cross-modal representations can effectively reduce the inter-modal gap, making it easier to locate specific persons using textual captions. Second, detailed correspondence is particularly important for TBPS, because fine-grained cross-modal information is often necessary to discriminate between similar persons [18].

In the TBPS domain, numerous approaches have been proposed from two distinct perspectives. First, some methods [19,20] employ only dual encoders and align the two modal representations using symmetric loss functions. Although computationally efficient, these

methods suffer from limited feature alignment capabilities. Progressive knowledge distillation also suggests that staged supervision can stabilize representation transfer in compact learning pipelines [21]. Alternative approaches [22] utilize multi-modal models with transformer-based [23] cross-attention mechanisms to improve cross-modal feature alignment and interaction, a goal related to diffusion-guided semantic consistency under multimodal heterogeneity [24]. However, such methods often require high computational costs due to the exhaustive fusion of all possible image-text pairs during inference, hindering their efficiency in real-time applications. Second, to obtain robust fine-grained features, recent TBPS works have designed multi-level [25–28], multi-granularity [29,30] matching strategies, and specialized attention modules [8,31,32]. Semantic prototype learning offers a related view that class-level semantic anchors can improve recognition stability when distributions shift [33]. While these methods provide fine-grained features, they suffer from complex model architectures and expensive computations. Furthermore, the fine-grained features they produce are often limited and hinder significant performance improvements.

To address these persistent data matching challenges, we propose a novel framework, SCMM, structurally optimized for calibrating cross-modal representations in TBPS. Our approach relies on a dual-encoder architecture tailored for high-speed indexing and retrieval computationally fit for large-scale databases. To fully explore inter-modal semantics without sacrificing efficiency, we propose two learning objectives. First, a new calibration loss steers alignment via quality-guided adaptive margins, selectively pushing apart negative instances and drawing in positive ones across modalities based on textual information density. Second, a Masked Caption Modeling (MCM) loss uncovers fine-grained word-level visual-textual connections. Operatively based on masked representation prediction, this phase activates a cross-modal decoder during training, thereby mapping fine-grained semantic connections without encumbering inference latency. Extensive experiments on three benchmarks clearly underscore our system’s scalability and robustness, yielding state-of-the-art performance across all tested environments. In summary, the primary contributions of this work are as follows:

- We propose SCMM, an efficient dual-encoder framework augmented with an auxiliary cross-modal decoder, strategically designed to overcome the classical trade-off between computational scalability and multimodal alignment precision for text-based person search.
- We introduce a new calibration loss anchored by quality-guided adaptive margins. By dynamically altering alignment boundaries based upon textual knowledge density, the loss diminishes the semantic gap and effectively organizes visual-textual knowledge representations.
- We formulate a masked caption modeling loss tailored to mine and formulate fine-grained cross-modal semantic correspondences, substantially heightening discriminatory capabilities among closely resembling instances.
- We validate the efficacy of SCMM through comprehensive experiments across CUHK-PEDES, ICFG-PEDES, and RSTPReID datasets, thereby validating the robustness of our method, with state-of-the-art Rank-1 accuracies corresponding to 73.81%, 64.25%, and 57.35%.

The remainder of this paper is organized as follows. Section 2 reviews related work on text-based person search and metric learning. Section 3 details the proposed SCMM framework, including the new calibration and masked caption modeling losses. Section 4 presents the experimental setup, comprehensive results on three benchmarks, ablation studies, and qualitative visualizations. Finally, Section 5 concludes the paper with a summary of contributions and future research directions.

2. Related work

2.1. Text-based person search

Text-based person search was first introduced by [8], which identifies person images in a gallery using only textual queries. Fundamentally, the task represents a critical application of multi-modal fusion, where the integration of heterogeneous data sources (visual and textual modalities) becomes essential for enhanced understanding and analysis capabilities. Multimodal medical surveys provide a related example in which heterogeneous visual and textual evidence must be fused carefully for dependable recognition [34]. Early works utilized pre-processing methods to obtain external cues such as person segmentation [25] and human body landmarks [35]. The evolution towards sophisticated fusion architectures has become crucial, as multi-view fusion combining complementary information from multiple sources shows significant promise for enhancing prediction accuracy and robustness in cross-modal tasks [12].

In recent years, end-to-end frameworks based on attention mechanisms [28,30,32] have become prevalent, addressing the fundamental challenge of semantic mismatch between features of different modalities. Cross-modal attention is critical for performing image-text interaction, where advanced fusion strategies must consider the unique attributes of cross-modal matching tasks [36]. Existing methods can be broadly classified into attention-explicit and attention-implicit approaches. Attention-explicit methods [6,32] design specific attention modules according to multi-granularity and multi-level strategies, implementing early fusion techniques that combine features at the input level. For example, NAFS [18] conducts cross-modal alignments over full-scale features with a contextual non-local attention module, while CFine [31] utilizes cross-attention for multi-grained global feature learning, achieving impressive results through knowledge transfer from the CLIP [37] model. In contrast, attention-implicit methods [28, 30] utilize transformer-based models with shared parameters to align cross-modal semantics implicitly, representing late fusion approaches that combine decisions from individual modality-specific models [12]. However, compared to performing in-depth cross-attention with task-driven approaches, existing methods face challenges in addressing information asymmetry problems and do not implement sufficient cross-interaction with generalized performance, highlighting the need for more sophisticated fusion techniques that enhance feature integration and establish deeper connections between global structure and local details [36]. To address these issues, various complementary strategies have emerged, such as serialized updating and matching [38], implicit local alignment [39], and cross-modal context sharing [40]. Furthermore, recent studies have explored robust feature learning under limited data constraints [41], and introduced local-enhanced representations [42] along with comprehensive attribute prediction [43] to capture fine-grained characteristics. More recent transformer-based TBPS systems, including IRRA [44], TP-TPS [13], and CLIP-enhanced variants such as CFine [31], have further strengthened retrieval performance through task-specific alignment designs or stronger pre-trained encoders. Their gains mainly come from richer feature interaction or better initialization. SCMM is positioned differently: it keeps the test-time dual-encoder retrieval path unchanged and sharpens supervision during training through quality-guided self calibration and masked caption modeling. The resulting novelty lies in jointly improving person-specific global alignment and token-level correspondence without introducing a heavier inference-time fusion pipeline.

2.2. Metric learning

Initially, metric learning used L_2 distance as the metric, with the goal of minimize the L_2 distance between samples of the same class. Some L_2 -based metric learning methods include Siamese Networks and

Triplet Networks [45]. With the development of deep learning, researchers started using softmax-based loss functions in metric learning to learn a more discriminative distance metric. Additionally, increasing the margin between classes is an intuitive approach to learn a better metric space [36]. L-Softmax introduced the concept of margin on the softmax function for the first time [35]. The widely used Cos-Face proposed large-margin cosine loss to learn highly discriminative deep features for face recognition [45]. Circle loss [46] proposed a simple loss based on a unified loss for metric learning and classification. Recently, some works introduced adaptive margin into marginal loss [6]. They usually learn image quality implicitly and adjust the margin accordingly. In single-modal representation learning, they usually give large margins to high-quality samples for hard mining. Globally optimized clustering for K-center objectives by Ren et al. [47] provides a strong and valuable complementary view of how compact structure can be effectively imposed on large-scale representations. Compared to them, we try to solve a cross-modal matching problem where samples from two modalities have different information volume. We give greater tolerance to less informative samples in TBPS. However, existing metric learning approaches face significant limitations when applied to cross-modal scenarios. First, traditional methods primarily focus on single-modal optimization and fail to adequately address the heterogeneity gap between different modalities []. Second, fixed margin constraints cannot adaptively handle the varying information density across modalities, leading to suboptimal alignment performance. Finally, most current approaches lack fine-grained correspondence modeling capabilities essential for distinguishing between similar cross-modal samples [3].

2.3. Masked language modeling

MLM is a highly effective method for pre-training language models [48] by randomly selecting a certain percentage of words from the input sentence and then predicting the masked words based on the context of other words. Many cross-modal pre-training models have utilized MLM in their methods. For example, the work in [22] combines MLM with contrastive loss in the framework, which achieved impressive performance in their cross-modal tasks. The success of MLM in BERT [49] has proven its ability to adapt well to various downstream tasks, leading to generalized performance. In our fine-grained framework, the task-driven decoder utilizing masked caption modeling can facilitate generic cross-modal learning. However, existing MLM-based cross-modal approaches face several limitations when applied to fine-grained person search tasks. First, traditional MLM methods primarily focus on intra-modal language understanding and lack explicit cross-modal interaction mechanisms, limiting their ability to establish detailed visual-textual correspondences [5]. Second, most existing approaches apply MLM uniformly without considering the varying information density and quality of textual descriptions, leading to suboptimal learning efficiency. Automated soft-sensor design offers an important insight into deployment-facing model construction by suggesting a related principle: model selection and adaptation should respond to signal quality rather than remain fixed across all inputs [50]. Survey work on diffusion-based anomaly detection further illustrates how generative objectives can strengthen discriminative learning when supervision is limited or imbalanced [51,52]. Finally, current methods often require complex multi-stage training procedures or heavy computational overhead during inference, hindering practical deployment in real-time applications [29].

3. Proposed methods

Formally, given a set of images with corresponding captions, denoted as $X = \{(I_i, T_i)\}_{i=1}^N$, where each image I_i and its description text T_i is associated with a person ID y_i , the text-based person search task aims to retrieve the most relevant Rank k (e.g., $k = 1, 5, 10$) person

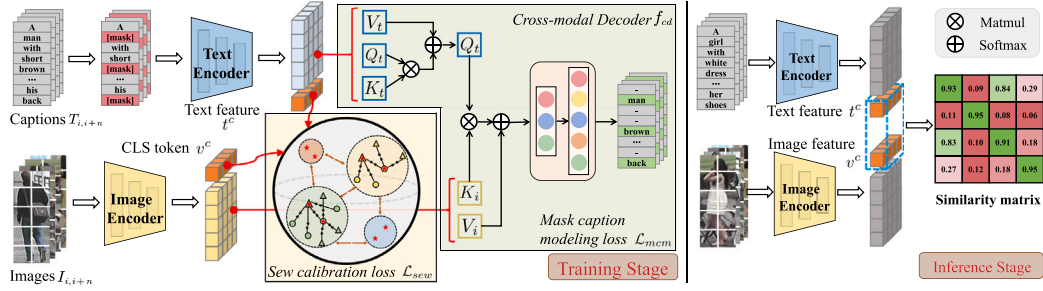


Fig. 2. Overview of our proposed SCMM. The framework consists of a dual-encoder for extracting image-text features and calibrating cross-modal representations with the sew calibration loss. We also include a decoder for performing cross-modal interaction with the task-driven Mask Caption Modeling. At the inference stage, we only utilize the classification (CLS) tokens from the dual-encoder to implement similarity search.

images efficiently from a large-scale gallery using textual queries. To accomplish this, the task requires learning discriminative cross-modal representations that can bridge the semantic gap between visual and textual modalities while maintaining fine-grained correspondence for accurate retrieval. The fundamental challenge lies in designing neural network architectures that can align heterogeneous feature representations, where visual features capture spatial and appearance characteristics while textual descriptions encode semantic attributes and contextual information. Effective learning strategies must address both global alignment across modalities and local correspondences that enable discrimination between visually similar individuals.

To address these challenges in cross-modal learning, we propose SCMM, a streamlined yet effective framework as illustrated in Fig. 2. Our approach employs a dual-encoder architecture using Vision Transformer (ViT) [53] and BERT [49] as backbone networks for visual and textual feature extraction, respectively. The image encoder processes image patches from I_i along with a vision classification (CLS) token, outputting an image feature sequence $\{v_i\}$ and a vision CLS token embedding v_i^c . Similarly, the text encoder generates a text feature sequence $\{t_i\}$ and a text CLS token embedding t_i^c for caption T_i , following established practices in vision-language learning [18,31]. Unlike CLIP-style pre-training [37], which learns a generic joint space through symmetric image-text contrast over web-scale pairs, SCMM preserves a TBPS-oriented dual-encoder backbone and injects cross-modal interaction only in the training phase. This design keeps retrieval efficient at inference time while retaining token-level textual representations required by masked caption modeling. Our framework integrates two complementary learning mechanisms to calibrate cross-modal representations: the sew calibration loss with adaptive margin constraints for global alignment, and the masked caption modeling loss for establishing fine-grained cross-modal correspondences. The sew calibration objective also differs from the CLIP pre-training loss because it couples bi-directional matching and identity-classification constraints with sample-wise adaptive margins, thereby emphasizing person-specific alignment rather than generic image-text agreement. The following subsections elaborate on each component. We first introduce the sew calibration loss that dynamically aligns cross-modal features based on textual information quality. Subsequently, we present the masked caption modeling mechanism that facilitates fine-grained correspondence learning through transformer-based cross-modal attention and masked prediction. The total loss function combines these complementary objectives to jointly optimize global cross-modal alignment and fine-grained correspondence learning, enabling the framework to achieve superior performance while maintaining computational efficiency through the streamlined dual-encoder architecture.

3.1. Sew calibration loss with constraints

In the TBPS task, closely aligning cross-modal representations is crucial for effectively finding specific persons with textual captions.

To address the heterogeneity between modalities, we propose a sew calibration loss that pushes each modality to a common space. As shown in Fig. 3(a), in single modality settings, a triplet distance constraint is widely utilized when embeddings are from a single modality distribution. In intra-class samples, the embeddings are pulled together, while inter-class samples are pushed away. In the cross-modal setting, we expect to impose constraints from both sides of the two embedding distributions. For example, in one-direction retrieval, we first need to align the features of a “perfect pair”, i.e., (I_i, T_i) . The shortest distance from the image embedding v_i to the text embedding should be t_i . Consequently, it is a perfect matching pair in person re-identification, and no other text feature will have a shorter distance (Eq. (1)). We set the perfect pair as an image-text anchor (A_{img}, A_{txt}) . Next, we impose another constraint between the image anchor A_{img} and its corresponding positive text samples $P_{txt}(\mathbb{1}(y_i = y_j, i \neq j))$ and negative text samples N_{txt} (Eq. (2)). For the other direction, we put symmetric constraints on A_{txt} and P_{img}, N_{img} . Fig. 3(b) shows the proposed constraints for cross modality. Forces from both sides act like a seam to pull the two distributions together.

Take the image side as an example, the L_2 distance constraints are shown as follows:

$$D(A_{img}, A_{txt}) + \mathcal{M}_1 < D(A_{img}, P_{txt}), \quad (1)$$

$$D(A_{img}, P_{txt}) + \mathcal{M}_2 < D(A_{img}, N_{txt}), \quad (2)$$

where D denotes L_2 distance. $\mathcal{M}_1$ and $\mathcal{M}_2$ are two margins, $\mathcal{M}_1 < \mathcal{M}_2$. With a bi-directional margin, each modal feature of the same pedestrian target is compressed compactly, making decision boundaries clearer. We then relax the constraints to:

$$0.5D^2(A_{img}, A_{txt}) + \mathcal{M}_1 < 0.5D^2(A_{img}, P_{txt}), \quad (3)$$

$$0.5D^2(A_{img}, P_{txt}) + \mathcal{M}_2 < 0.5D^2(A_{img}, N_{txt}). \quad (4)$$

Subsequently, we change the above pairwise constraints into soft forms for better convergence following [46].

$$\mathcal{L}_{match}^{Pull} = \log \left[1 + \sum_{k=1}^K \exp(\alpha(\bar{v}_i^c \bar{t}_k^c - \bar{v}_i^c \bar{t}_i^c + \mathcal{M}_1)) \right], \quad (5)$$

$$\mathcal{L}_{match}^{Push} = \log \left[1 + \sum_{k=1}^K \sum_{j=1}^J \exp(\alpha(\bar{v}_i^c \bar{t}_j^c - \bar{v}_i^c \bar{t}_k^c + \mathcal{M}_2)) \right], \quad (6)$$

where $\bar{v}^c$ and $\bar{t}^c$ are CLS token features after normalizing, K and J denote the number of positive and negative samples in this batch, respectively. α is a scale parameter. As a result, the image-to-text matching part of the sew calibration loss is formulated as below:

$$\mathcal{L}_{match}^{I2T} = \mathcal{L}_{match}^{Pull} + \mathcal{L}_{match}^{Push}. \quad (7)$$

The constraints in Eq. (2) can also be used to impose a classification loss in a similar way. As there is no difference for perfect positive samples

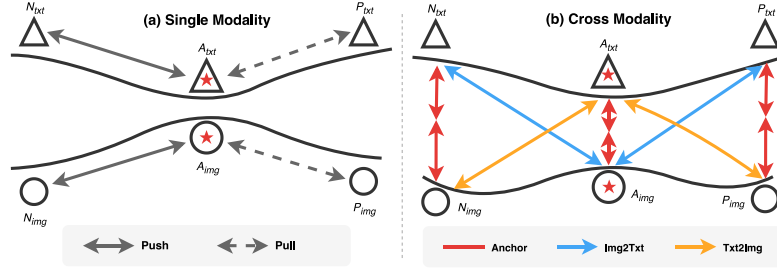


Fig. 3. Illustration of sew calibration loss. The constraints are different between single-modal and cross-modal matching. (A_{img}, A_{txt}) denotes anchors in image-text feature distribution, while (P_{img}, P_{txt}) and (N_{img}, N_{txt}) denote positive and negative sample pairs, respectively. The sew calibration loss pushes negative sample pairs and pulls positive sample pairs, stitching cross-modal key information like a seam.

in the classification task, we omit to constrain Eq. (1). Formally, the loss for our person ID classification part is as follows [20]:

$$\mathcal{L}_{id}^{I2T} = \frac{1}{n} \sum_{i=1}^n -\log \left(\frac{e^{(\alpha(s_{y_i,i} - \mathcal{M}_2))}}{e^{(\alpha(s_{y_i,i} - \mathcal{M}_2))} + \sum_{j \neq y_i} e^{(\alpha(s_{y_j,i} - \mathcal{M}_2))}} \right), \quad (8)$$

$$s_{i,j} = \omega_i^T \hat{v}_j, \quad \hat{v}_i = (v_i^c)^T \tilde{r}_i^c \cdot \tilde{r}_i^c, \quad (9)$$

where n is batch size, and ω represents the classification weight after normalization. $\hat{v}_i$ can be explained as the projection of image representation v_i^{cls} onto the normalized text representation $\tilde{r}_i^c$. Intuitively, $\hat{v}_i$ keeps only the component of the image CLS representation that lies along the paired text direction. The ID classifier therefore scores whether the image embedding is projected onto the correct text-conditioned identity axis, which makes Eqs. (8) and (9) easier to interpret: the margin term encourages the projected image feature to stay closer to its matched identity than to competing identities in the shared space.

$\mathcal{L}_{match}^{T2I}$ and $\mathcal{L}_{id}^{T2I}$ are in the same form as above, but the change is focused on the text-to-image. Both matching and classification loss have identical decision boundaries. Equipped with our proposed cross-modal constraints, the sew calibration loss can effectively reduce the gap between image and text feature distributions. Although the margin restrictions allow our model to learn better cross-modal representations, using a fixed margin $\mathcal{M}$ in all cases may not be flexible enough. A large margin constraint makes model learning difficult, while too small a margin does not impose a significant constraint. An adaptive margin guided by quality can be more effective. In TBPS, the texts come from the annotator's descriptions of the images. Images are expected to have complete information, while texts have varying amounts of information. Thus, we adjust the margin value based on the quality of the text description. Token length is used here as a coarse proxy for textual information density rather than a direct measure of semantic usefulness: a short caption may still contain highly discriminative attributes, whereas a longer description can include redundancy or exceed the text encoder's effective context usage. The adaptive margin therefore serves as a monotonic prior on annotation richness, giving looser constraints to less detailed captions while preserving a simple formulation that is stable across datasets. Based on this design, we compute the adaptive margin for each image-text pair according to its text total tokens length $\mathcal{T}_i$:

$$\mathcal{M}_i = \mathcal{M}_{min} + \frac{(\mathcal{M}_{max} - \mathcal{M}_{min}) \cdot (\mathcal{T}_i - \mathcal{T}_{min})}{\mathcal{T}_{max} - \mathcal{T}_{min}}, \quad (10)$$

where $\mathcal{M}_{max}$ and $\mathcal{M}_{min}$ are upper and lower bounds of margins. $\mathcal{T}_{max}$ and $\mathcal{T}_{min}$ are bounds of the captions length, respectively. We set $\mathcal{T}_{max}$ and $\mathcal{T}_{min}$ according to the different dataset captions length distributions. After that, we utilize $\mathcal{M}_i$ to replace the fixed margin $\mathcal{M}$ above. We simply set $\mathcal{M}_1 = \mathcal{M}_2 = \mathcal{M}_i$. Each $\mathcal{M}_i$ is computed independently from the token count of caption $\mathcal{T}_i$ before the batch losses are aggregated. The normalization uses the dataset-level bounds $(\mathcal{T}_{min}, \mathcal{T}_{max})$ rather than the current mini-batch statistics, so the adaptive margin is not re-scaled by the composition of a particular batch.

Algorithm 1: Training Procedure of SCMM

Input: Image I and text T ; a batch of n paired samples $\mathbb{B} = \{(I^1, T^1), (I^2, T^2), \dots, (I^n, T^n)\}$

Output: Training loss $(\mathcal{L}_{sew}, \mathcal{L}_{mcm})$ or embeddings (I_{cls}, T_{cls})

for each T^i **in** $\mathbb{B}$ **do**

$T^i \leftarrow \text{mask}(T^i)$

end for

for $i \leftarrow 1$ **to** n **do**

$(I_{cls}^i, T_{cls}^i, \dots, I_c^i) \leftarrow ViT(I^i)$

$(T_{cls}^i, T_1^i, \dots, T_c^i) \leftarrow Bert(T^i)$

if Training Stage **then**

$(I_{cls}^i, T_{cls}^i) \leftarrow BatchNorm(I_{cls}^i, T_{cls}^i)$

$I_{context}^i, T_{context}^i \leftarrow (I_1^i, \dots, I_c^i), (T_1^i, \dots, T_c^i)$

$T_{context}^i \leftarrow Attn(T_{context}^i)$

$T_{cross}^i \leftarrow CrossAttn(T_{context}^i, I_{context}^i)$

$\mathcal{L}_{sew}^i \leftarrow SewCalibration(I_{cls}^i, T_{cls}^i)$

$\mathcal{L}_{mcm}^i \leftarrow MCM(T_{cross}^i, T_{cls}^i)$

else

$\{(I_{cls}^1, T_{cls}^1), (I_{cls}^2, T_{cls}^2), \dots, (I_{cls}^i, T_{cls}^i)\} \leftarrow (I_{cls}, T_{cls})$

end if

end for

if Training Stage **then**

return $(\mathcal{L}_{sew}, \mathcal{L}_{mcm})$

else

return (I_{cls}, T_{cls})

end if

3.2. Masked caption modeling loss

TBPS is a fine-grained cross-modal task, which means that only caption-level discrimination is not enough. If the textual captions of two persons differ in a few words, a TBPS method cannot retrieve a specific person without word-level discrimination. Although there are many works to establish word-level discrimination capacities, such methods are complex and limited, hindering performance boost. To solve this issue, we propose masked caption modeling to establish detailed image-text relationships. Furthermore, by utilizing MCM, our framework can perform more generic cross-modal learning.

Inspired by [5, 48], we add a masked prediction task on the text branch. Concretely, we mask a portion of text tokens (replacing them with a learnable token vector) and employ a cross-modal decoder f_{cd} . The text encoder outputs the corresponding text features, and the decoder maximizes the conditional likelihood of the masked text feature t_n given the image feature sequence $\{v_i\}$ and text feature sequence $\{t_i\}$.

$$\mathcal{L}_{mcm} = - \sum_{n=1}^N \log P_{\theta}(t_n | \{t_i\}, \{v_i\}), \quad (11)$$

where N is the total masked tokens numbers in a caption.

As shown in Fig. 2, the cross-modal decoder part, the decoder f_{cd} takes both unmasked text tokens and masked tokens in their original order as the input. The multi-head self-attention [23] first encodes the text features as Q_t, K_t, V_t , while the cross self-attention further improves the text features by taking into account the encoded image features as K_i, V_i for visual context. The final linear projection layer has the same number of output channels as the text vocabulary and computes the cross-entropy loss between the reconstructed and original words only on masked text tokens. It should be noted that f_{cd} is only used during training and not during inference.

3.3. Total loss

The total loss function of SCMM integrates our two complementary learning mechanisms to achieve comprehensive cross-modal alignment and correspondence learning. Our unified optimization objective combines the sew calibration loss and the masked caption modeling loss to jointly address both global alignment and fine-grained correspondence challenges in cross-modal representation learning. The sew calibration loss $\mathcal{L}_{sew}$ includes bidirectional matching and classification components that ensure symmetric cross-modal alignment:

$$\mathcal{L}_{sew} = \mathcal{L}_{match}^{I2T} + \mathcal{L}_{match}^{T2I} + \mathcal{L}_{id}^{I2T} + \mathcal{L}_{id}^{T2I}, \quad (12)$$

where the image-to-text and text-to-image matching losses establish cross-modal correspondence constraints, while the identification losses enforce discriminative learning for person-specific features. Importantly, the bidirectional formulation ensures that both modalities are effectively aligned in the shared embedding space through symmetric constraints. Subsequently, the complete loss function combines both components through weighted summation:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{sew} + \lambda_2 \mathcal{L}_{mcm}, \quad (13)$$

where λ_1 and λ_2 are hyperparameters balancing global alignment and fine-grained correspondence learning. Through this joint optimization, the framework simultaneously learns compact cross-modal representations while capturing detailed visual-textual relationships for discriminating similar individuals. The complementary loss components create synergistic effects ensuring comprehensive representation learning, addressing both inter-modal gap reduction and fine-grained correspondence modeling. Regarding computational complexity, SCMM maintains efficient processing through its streamlined dual-encoder architecture.

4. Experiments

4.1. Datasets and evaluation metric

We evaluate SCMM on three benchmarks: CUHK-PEDES [8], ICFG-PEDES [26], and RSTPReID [17]. As the first large-scale benchmark for text-based person search tasks, CUHK-PEDES contains 40,206 images of 13,003 person IDs collected from five person re-identification datasets, with each image having two different textual captions with an extensive vocabulary and an average sentence length of 23.5 words. Additionally, the testing set comprises 3074 images and 6148 descriptions of 1000 persons. ICFG-PEDES contains 54,522 images of 4102 persons, where each image’s corresponding description has an average length of 37 words, and the testing set consists of 19,848 image-text pairs of 1000 persons. RSTPReID contains 20,505 images of 4101 persons, with each pedestrian having five images. It is divided into a training set with 3701 persons, a validation set with 200 persons, and a testing set with 200 persons. For our evaluation metric, we report the Rank k ($k = 1, 5, 10$) text-to-image accuracy, which is commonly used in previous works to evaluate text-based person search. Given a textual description as the query, if the top- k retrieved images contain any person corresponding to the query, we consider it a successful person search.

Table 1

Comparison with SOTA methods on the CUHK-PEDES dataset. Rank1, Rank5, and Rank10 accuracies (%) are reported. Bold number represents the best score. R and C denote the re-ranking post-processing operations and image encoder pre-trained on CLIP model.

Method	Rank1	Rank5	Rank10
GNA-RNN [8]	19.05	–	53.64
CMPM/C [20]	49.37	–	79.27
Dual-path [19]	44.40	66.26	75.07
ViTAA [25]	55.97	75.84	83.52
VP Net [35]	58.83	81.25	86.72
SUM [38]	59.22	80.35	87.60
CLIP [37]	60.67	81.99	88.87
DSSL [17]	59.98	80.41	87.56
MGEL [29]	60.27	80.01	86.74
SSAN [26]	61.37	80.15	86.73
TestReID [41]	64.08	81.73	88.19
ACSA [32]	63.56	81.40	87.70
IVT [53]	64.00	82.72	88.95
SAFA Net [28]	64.13	82.62	88.40
TIPCB [27]	64.26	83.19	89.10
CAIBC [14]	64.43	82.87	88.37
AXM Net [40]	64.44	80.52	86.77
CFine [31]	65.07	83.01	89.00
ISANet [39]	63.92	82.15	87.69
CAPL [43]	65.63	84.81	90.21
LERF [42]	65.84	84.24	90.22
SCMM	67.71	84.57	89.44
SCMM+R	71.09	86.78	91.23
SCMM+C	69.61	86.01	90.90
SCMM+R+C	73.81	88.89	92.77

4.2. Implementation details

In our visual-textual dual-encoder, we extract visual representations using the ViT-Base pre-trained on ImageNet [54]. The images are resized to 224×224 pixels. For textual representations, we use the BERT-Base-Uncased model pre-trained on the Toronto Book Corpus and Wikipedia. The representation dimension is set to 768, and the feature sequence lengths are set to 197 and 100, respectively. During the training phase, we use a batch size of 64 and train for 60 epochs. We use Adam optimizer with an initial learning rate of 0.001. To augment our data, we apply a random horizontal flipping operation, and we use a mask ratio of 0.1 for randomly masking text tokens. After selecting this value on CUHK-PEDES through the ablation in Table 7, we keep the same 0.1 mask ratio for ICFG-PEDES and RSTPReID rather than tuning a separate ratio for each dataset. Only the caption-length bounds ($\mathcal{T}_{min}, \mathcal{T}_{max}$) are adjusted to match dataset-specific caption distributions. The minimum and maximum textual information length boundaries $\mathcal{T}_{min}$ and $\mathcal{T}_{max}$ are set to 20–60, 25–65, and 22–60 according to the caption length distributions of CUHK-PEDES, ICFG-PEDES, and RSTPReID, respectively. The bounds of the upper and lower margin $\mathcal{M}$ are set to 0.4 and 0.6, and the scale parameter α is set to 32. For each loss in the total loss function, the balance factors λ_1 and λ_2 are set equal to 1. During the testing phase, we apply a re-ranking post-processing approach to improve search performance following NAFS [18]. We conduct the experiments on four NVIDIA 2080Ti with PyTorch.

4.3. Comparison with state-of-the-art methods

4.3.1. Results on CUHK-PEDES

Table 1 compares our framework and previous methods on CUHK-PEDES. To keep the comparison fair, it is useful to read the results in three groups: standard methods without re-ranking or CLIP-enhanced initialization, methods with re-ranking, and variants that additionally import a CLIP-pre-trained visual encoder. Under the standard

Table 2
Quantitative results on the ICFG-PEDES dataset.

Method	Rank1	Rank5	Rank10
CMPM/C [20]	43.51	65.44	74.26
Dual-path [19]	38.99	59.44	68.41
VITAA [25]	50.98	68.79	75.78
SSAN [26]	54.23	72.63	79.53
CLIP [37]	53.96	73.69	80.43
TIPCB [27]	54.96	74.72	81.89
IVT [30]	56.04	73.60	80.22
CFine [31]	55.69	72.72	79.46
CFine+C [31]	60.83	76.55	82.42
IRRA [44]	63.46	80.25	85.82
TP-TPS [13]	60.64	75.97	81.76
ISANet [39]	57.73	75.42	81.72
LERF [42]	57.23	76.64	83.11
CANC [6]	60.52	78.36	84.13
SCMM	60.20	75.97	81.78
SCMM+R	62.17	85.74	89.67
SCMM+C	62.29	77.15	82.52
SCMM+R+C	64.25	86.95	90.70

Table 3
Quantitative results on the RSTPReid dataset.

Method	Rank1	Rank5	Rank10
CLIP [37]	50.10	76.10	84.95
DSSL [17]	32.43	55.08	63.19
SUM [38]	41.38	67.48	76.48
SSAN [26]	43.50	67.80	77.15
IVT [30]	46.70	70.00	78.80
ACSA [32]	48.40	71.85	81.45
LERF [42]	46.75	71.30	81.60
CFine [31]	45.85	70.30	78.40
CFine+C [31]	50.55	72.50	81.60
TP-TPS [13]	50.65	72.45	81.20
SCMM	50.75	74.20	81.70
SCMM+R	55.35	77.30	84.25
SCMM+C	51.95	73.50	82.45
SCMM+R+C	57.35	77.50	85.50

setting, our proposed SCMM outperforms the previous best method CFine [31] by 2.64%, 1.56%, and 0.44% on Rank1, Rank5, and Rank10, respectively. The state-of-the-art results on CUHK-PEDES show the effectiveness of SCMM in the same inference regime. Re-ranking then provides an additional boost, raising Rank1 to 71.09%. Furthermore, with a stronger image encoder pre-training on CLIP, SCMM achieves 73.81%, 88.89%, and 92.77% on the three metrics, respectively. In this comparison, SCMM + C replaces only the visual backbone with CLIP-pre-trained initialization [37] while retaining the BERT-based text encoder [49]. That configuration isolates the effect of transferring a stronger visual encoder into the proposed framework and keeps the masked caption modeling branch compatible with token-level word prediction. Replacing the text encoder with a CLIP text tower is also meaningful, but it would require a joint redesign of the textual token interface and the training-only decoder rather than an isolated backbone swap. Overall, the grouped comparison shows that SCMM scales consistently across the base setting, stronger visual pre-training, and extra post-processing operations.

4.3.2. Results on ICFG-PEDES and RSTPReid

We also utilized other benchmarks to validate SCMM’s performance and generalization. The ICFG-PEDES and RSTPReid datasets are more challenging compared to CUHK-PEDES, and our method significantly outperformed all state-of-the-art methods on these two datasets by a large margin, as reported in Tables 2 and 3. Compared with the state-of-the-art results [31] on ICFG-PEDES, SCMM achieved significant

Table 4
Performance comparisons of various components.

ID	Res	ViT	Sew-F	Sew-A	MCM	Rank1	Rank5	Rank10
1	✓					60.39	80.22	86.53
2		✓				62.29	81.25	87.22
3		✓	✓			65.41	82.56	88.01
4		✓		✓		66.02	83.33	88.61
5		✓			✓	65.16	81.95	87.93
6	✓		✓	✓	✓	67.71	84.57	89.44

Notes: Res and ViT denote ResNet and Vision Transformer as image encoders, respectively. Sew-F and Sew-A refer to the fixed and adaptive margin versions of the sew calibration loss, respectively. Model 2 is our baseline.

improvements, with scores of 60.20% (+4.51%), 75.97% (+3.25%), and 81.78% (+2.32%) on Rank1, Rank5, and Rank10, respectively. On the RSTPReid dataset, we achieved scores of 50.75% (+4.90%), 74.20% (+3.90%), and 81.70% (+3.30%) on the three metrics. With re-ranking post-processing, we were able to achieve scores of 62.17% and 55.35% on Rank1 for the two benchmarks, respectively. The gain on RSTPReid is more modest before re-ranking: SCMM improves Rank1 from 50.65% with TP-TPS [13] to 50.75%, which is still a positive but narrower margin than the headline results obtained with re-ranking or CLIP-enhanced initialization. A plausible reason is that RSTPReid contains fewer testing identities and many visually similar pedestrians [17], making fine-grained distinctions harder when retrieval relies only on the base embedding geometry. Re-ranking should therefore be interpreted as a separate post-processing enhancement rather than part of the core base-model comparison, and it brings a significant boost because the learned embeddings already preserve compact cross-modal neighbor relations. Moreover, by utilizing the CLIP pre-trained image model as an encoder, we achieved even better results, with scores of 64.25%, 86.95%, and 90.70% on ICFG-PEDES, and scores of 57.35%, 77.50%, and 85.50% on RSTPReid. Consequently, the two challenging benchmark results demonstrate the robustness and scalability of SCMM across the base setting and the stronger CLIP-initialized variant.

4.4. Ablation studies

4.4.1. Analysis of model components

To validate SCMM’s components, we evaluate their contributions on CUHK-PEDES (Table 4). **Models 1** and **2** compare ResNet-50 and ViT-Base encoders, with ViT showing superior performance (+1.90% Rank1). Using Model 2 as the baseline, **Models 3** and **4** demonstrate that the sew calibration loss improves alignment, with adaptive margins (**Model 4**) outperforming fixed margins (**Model 3**). **Model 5** shows that MCM significantly boosts performance by mining fine-grained details. Notably, the full SCMM framework (**Model 6**) combines these components to achieve the best results, outperforming the baseline by 5.42% in Rank1 accuracy. The separate gains from Sew-A (+3.73%) and MCM (+2.87%) are therefore not strictly additive, indicating partial overlap in the errors they correct. At the same time, the full model still surpasses either single component by 1.69%–2.55%, which suggests a positive interaction: sew calibration improves the global embedding geometry, while MCM contributes complementary token-level cues that are more useful once the two modalities are already better aligned. The consistent improvement confirms that jointly optimizing global alignment and fine-grained correspondence is essential for effective cross-modal representation learning.

4.4.2. Impact of sew calibration loss for reducing cross-modal gap

Benefiting from the sew calibration loss, which reduces the cross-modal gap, the learned representation distributions are closer than with the basic loss. To demonstrate this, we illustrate a comparison of singular value decomposition on cross-modal representations [36].

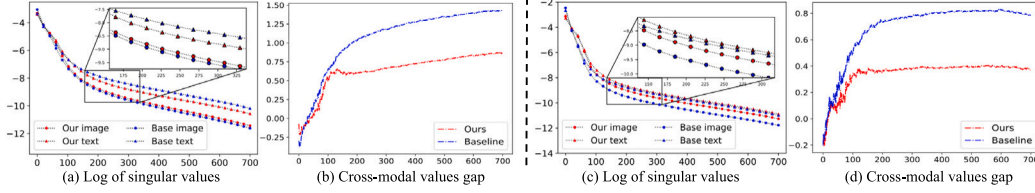


Fig. 4. Comparison of singular values for image-text embedding features across CUHK-PEDES (a-b) and ICFG-PEDES (c-d) datasets. (a) and (c) depict the distribution of singular values, where a smaller inter-line gap indicates a closer cross-modal distribution. (b) and (d) present the logarithmic difference in singular values between the baseline and our approach for CUHK-PEDES and ICFG-PEDES, respectively.

Table 5

MCM components on the CUHK-PEDES dataset.

Mask	Attention	Caption modeling	Rank1	Rank5	Rank10
✓			63.65	81.84	87.51
	✓		63.83	80.87	86.35
✓		✓	64.28	81.97	87.75
✓	✓	✓	65.16	81.95	87.93

Specifically, Fig. 4(a) and (c) present the baseline and SCMM’s singular value decomposition for the text and image modalities. We computed the distance between these two modalities at specific singular values and reflected them in Fig. 4(b) and (d). Intuitively, the smaller the distance between different modality features, the closer the distribution learned by the model. We find that the sew calibration loss performs better representation distributions in the common embedding space. It ensures closer cross-modal representation distributions and a smaller inter-modal gap than the baseline.

4.4.3. Impact of masked caption modeling

The masked caption modeling operation in fine-grained cross-modal interaction is critical in our framework. It comprises token masking, an attention module, and masked tokens prediction. As shown in Table 5, we explore the effectiveness of MCM components on the CUHK-PEDES. First, only masking on the input text tokens without a reconstruction task behaves like a random erase text augmentation. Remarkably, the augmentation can already bring +1.36% marginal improvement and reach 63.65% on Rank1. It shows that the details provided by word tokens are helpful for fine-grained recognition. Next, if we only utilize the attention mechanism to enhance cross-modal representations without the mask caption modeling, it can achieve 63.83% on Rank1. We explain this improvement as the cross-attention brings details from sequence tokens to the CLS token. Meanwhile, image features also contribute a lot to this process.

On the other hand, we also design an experiment of mask caption modeling without the cross attention. The decoder f_{cd} directly predicts all masked tokens from dual-encoder outputs. In this experiment, we observe 64.28% on Rank1. It proves the reconstruction tasks in the decoder help guide the text encoder to learn richer and more refined representations. We also notice that with all three components, the MCM achieves 65.16% on Rank1. With the help of cross attention from image features, the CLS token ensembles all those rich information.

4.4.4. Analysis of generalisability validation

To better understand the contributions of MCM in our framework, we conduct a domain generalization analysis on the three benchmarks to demonstrate our generalization performance. In Table 6, CUHK $\Rightarrow$ ICFG and CUHK $\Rightarrow$ RSTP indicate using the model trained on the CUHK-PEDES dataset to infer their test sets and vice versa. Compared to the baseline, we can observe a performance improvement in the three metrics for ICFG-PEDES and RSTPReid. The fine-grained features mined by MCM are more resistant to overfitting. People-centric domain generalization provides related evidence that human-centered representation

Table 6

Performance analysis for cross-domain validation.

CUHK $\Rightarrow$ ICFG	Rank1	Rank5	Rank10
Baseline	30.79	49.10	57.47
Baseline+MCM	31.13	49.52	58.07
ICFG $\Rightarrow$ CUHK	Rank1	Rank5	Rank10
Baseline	22.19	40.55	50.52
Baseline+MCM	23.78	42.41	52.11
CUHK $\Rightarrow$ RSTP	Rank1	Rank5	Rank10
Baseline	37.40	63.40	74.35
Baseline+MCM	39.25	63.95	74.55
RSTP $\Rightarrow$ CUHK	Rank1	Rank5	Rank10
Baseline	10.02	22.95	31.04
Baseline+MCM	10.69	25.20	33.93

models should maintain discriminative behavior when data distributions shift across domains [55]. The improvement in domain generalization shows the capability of our framework in learning generic cross-modal information, which is essential to solving the fine-grained modal heterogeneity.

4.4.5. Effect of margin parameters

In Section 3.1, we introduce the adaptive margins to constrain cross-modal learning using Sew Calibration loss with constraints. We find that compared to the manually fixed parameters (SEW-f), our adaptive margins (SEW-a) produce better results, as shown in Fig. 5, where we investigate the effect of manual margin in different settings on the CUHK-PEDES [8] on Rank1 accuracy. The best performance of 65.41% is achieved when $\mathcal{M} = 0.5$, while the worst model performance of 62.29% is observed when $\mathcal{M} = 0$. We note that the performance first improves as the margin increases, which is due to our bi-directional margin compressing the same pedestrian target compactly. However, we find that the performance does not continue to improve as the margin increases, and we observe an inflection point around a margin of 0.5. When the margin is too large, the tight constraints make the model too hard to train. Therefore, it is critical to utilize quality-guided adaptive margins according to the textual information. In this regard, we set the upper and lower margin bounds ($\mathcal{M}_{min} = 0.4$ and $\mathcal{M}_{max} = 0.6$). In this range, our Sew Calibration loss with adaptive constraints obtains the best performance, which a manually fixed margin cannot achieve, regardless of how the margin value is tuned.

4.4.6. Effect of mask ratio

Inspired by Masked Language Modeling [49], we propose the MCM loss which leverages a masked captions prediction task to establish detailed and generic relationships between textual and visual parts. Our method requires setting the mask ratio in the MCM. To investigate the effect of the mask ratio on model performance and generalization learning ability, we conducted ablation experiments, and the results are shown in Table 7. When the mask ratio is set to 0, which is equivalent to the baseline, the performance is the worst. Compared to

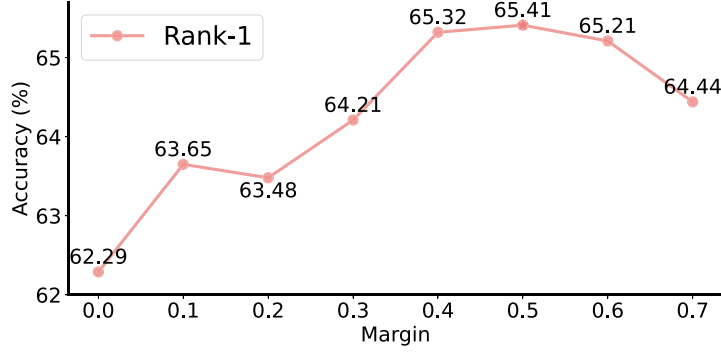


Fig. 5. Effect of the manual fixed margin parameter settings of our Sew Calibration loss in terms of Rank1 accuracy on the CUHK-PEDES.

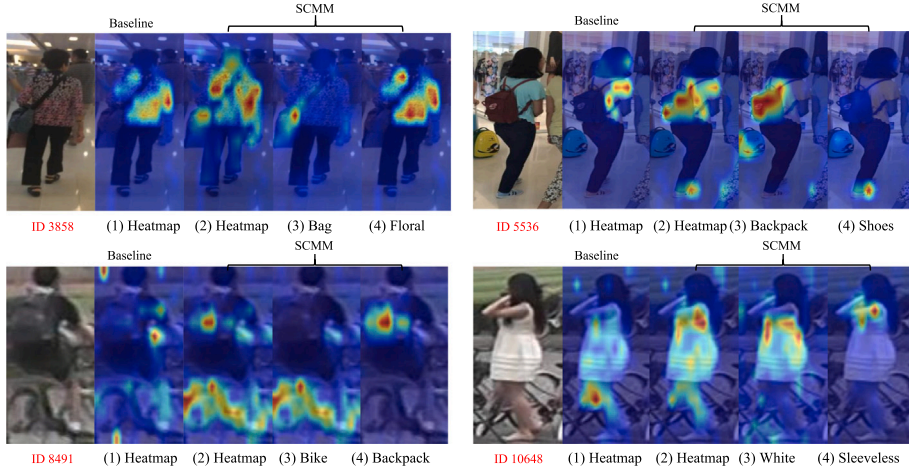


Fig. 6. Visualization of attention maps from baseline and SCMM on the CUHK-PEDES. We present the total caption-level results and the fine-grained word-level results, respectively. The caption-level heat maps average the last-layer attention over heads, while each word-level map uses the selected token as the query and averages its last-layer cross-attention over heads without averaging across different keywords. (1) Baseline caption-level results, (2) Our caption-level results, (3) & (4) Our word-level results.

Table 7

Ablation results of mask ratio on the CUHK-PEDES.

Baseline+MCM	Rank1	Rank5	Rank10
Mask = 0	62.29	81.25	87.22
Mask = 0.1	65.16	81.95	87.93
Mask = 0.3	64.97	81.38	87.71
Mask = 0.5	63.78	81.11	87.43

the baseline, our method with a 0.1 mask ratio achieves the best results, with scores of 65.16%, 81.95%, and 87.93% on Rank1, Rank5, and Rank10, respectively. The experiments are conducted on the CUHK-PEDES dataset using the Baseline + MCM method. As shown in these results, the model achieves the best performance when the mask ratio is set to 0.1, and the performance gradually decreases as the ratio increases. We therefore keep 0.1 as a single shared hyperparameter on ICFG-PEDES and RSTPreID instead of re-tuning it per dataset, so the cross-dataset comparisons reflect the same MCM configuration. The main reason is that the information provided in the annotated caption text is already accurate and relatively concise. As the mask ratio expands, more key information may be obscured, causing the model to fail to learn enough semantic information to complete the construction of the image-text relationship.

4.5. Qualitative analysis

4.5.1. Visualization of attention map

We visualize attention maps on the CUHK-PEDES test dataset to demonstrate model capability in learning image-text correspondence. As shown in Fig. 6, the visualization obtained by SCMM is more apparent and refined than the baseline. For the caption-level panels, the displayed response summarizes the final-layer attention pattern across heads. For the word-level panels, each heat map is generated from one selected keyword in the caption description, such as bag, shoes, colors, and clothing, and no additional averaging is performed across different tokens. For example, the visualizations of IDs 8491 and 10648 do not focus on the most useful detailed information under the baseline. The key messages in the two images are a man riding a bike with a backpack and a woman wearing a sleeveless white dress. SCMM captures these key details more reliably, showing that the MCM branch learns critical parts in the cross-modal correspondence.

4.5.2. Visualization of image-text feature distribution

SCMM is capable of learning better image-text representation distributions with fine-grained image-text interaction and adaptive constraints to reduce the cross-modal gap. To demonstrate this, we randomly selected some pedestrians and extracted their image-text global features, then mapped them to two dimensions for visualization in Fig. 7(a) and (b). For instance, we can take IDs 5532 and 11800 as examples. We can observe that the image-text features of the same

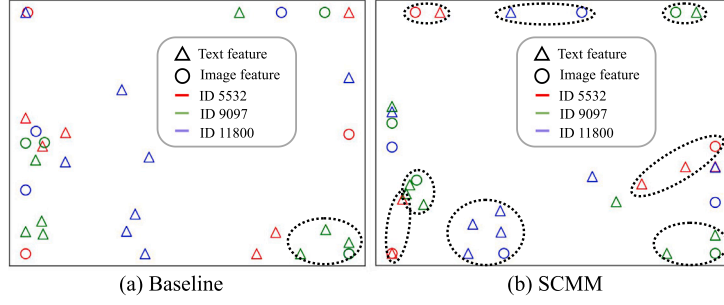


Fig. 7. Presentation of cross-modal representation distributions on CUHK-PEDES test dataset. Different colors correspond to the different target ID.



Fig. 8. Top-5 retrieval results on CUHK-PEDES. Matched target person images are marked with red rectangles. The subcaption below each image shows the text query used during retrieval.

ID distribution are compressed more compactly, and the boundaries between them are more apparent than the baseline results, which demonstrates the effectiveness of SCMM in mitigating the cross-modal gap.

4.5.3. Visualization of top-5 retrieval results

To comprehensively evaluate the retrieval performance of our method, we randomly selected several sets of search results and visualized them in Fig. 8. For each target image, we show the top-5 most similar search results, with similarity scores after re-ranking displayed above each image. Correct retrievals are indicated by red rectangles, and the number in the last column shows how many pedestrian targets sharing an identical ID have been recalled.

Taking ID 5219 as an example, our method correctly retrieved all three target images, with similarity scores of 1.2563, 1.2100, and 0.9775, respectively. The two unmatched images received substantially lower similarity scores of 0.5174 and 0.5011. Although those images share partial visual similarities with the positive samples, they conflict with key details in the text description, demonstrating that SCMM has learned to differentiate subtle distinctions. The final row of Fig. 8 illustrates a particularly challenging case where the query person possesses

highly similar attributes to multiple other pedestrians; nevertheless, SCMM correctly ranks two out of three positive images as the top-2 results.

5. Conclusion

In this paper, we proposed SCMM, an innovative multimodal perception framework designed to calibrate cross-modal representations for text-based person search, advancing multimodal pattern recognition research. Focusing on the precise alignment of visual and textual semantics, SCMM uniquely integrates a new calibration loss anchored by quality-guided adaptive margins, establishing global semantic alignment, and a masked caption modeling loss that facilitates complex fine-grained concept correspondence. Crucially, our dual-encoder architecture employs a training-only cross-modal decoder, thereby avoiding computational overhead during inference while maximizing cross-modal pattern discovery capabilities. Extensive experiments across three large-scale benchmark datasets consistently demonstrate that SCMM significantly advances the state-of-the-art framework boundaries, achieving 73.81% Rank-1 retrieval accuracy on CUHK-PEDES, validating the effectiveness of our approach for multimodal perception and visual pattern recognition.

Despite these strengths, several limitations merit acknowledgment. First, the quality-guided adaptive margin currently uses text token length as a proxy for informational density, so it cannot explicitly distinguish concise but highly discriminative captions from long descriptions containing redundant phrases. Richer quality signals based on semantic complexity, attribute diversity, or encoder confidence may yield further gains. Second, SCMM requires fully annotated image-text pairs during training, which limits scalability to weakly supervised or cross-lingual retrieval scenarios where matched pairs are scarce. Third, the framework's discriminative capacity may diminish under severe occlusion or drastic appearance changes across non-overlapping camera views, an open challenge shared across TBPS methods. The broader pattern recognition value of SCMM lies in its modular and efficient alignment design: the dual-encoder backbone remains suitable for large-scale retrieval, while sew calibration and MCM can be adapted to related cross-modal frameworks without adding inference-time cost. The domain-generalization results in Table 6 further indicate that MCM learns transferable cross-modal information across CUHK, ICFG, and RSTP. Future work will therefore focus on richer semantic quality estimators beyond token length, weaker supervision for scarce paired data, and extensions to video-based or open-vocabulary person retrieval.

CRediT authorship contribution statement

Jing Liu: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Donglai Wei:** Writing – original draft, Methodology, Data curation, Conceptualization. **Yang Liu:** Writing – review & editing, Methodology, Conceptualization. **Sipeng Zhang:** Writing – review & editing, Funding acquisition. **Tong Yang:** Writing – review & editing, Funding acquisition. **Wei Zhou:** Writing – review & editing, Conceptualization. **Weiping Ding:** Writing – review & editing, Conceptualization. **Victor C.M. Leung:** Writing – review & editing, Supervision, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used generative AI and AI-assisted technologies for writing enhancement and language polishing to improve the clarity and readability of the manuscript. After using these tools, the authors have reviewed and edited the content as needed and take full responsibility for the content of the publication.

Declaration of competing interest

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

Acknowledgments

This work was supported in part by the Mitacs Elevate under Grant No. IT44479, Canada; the Guangdong Pearl River Talents Recruitment Program under Grant No. 2019ZT08X603, China; and the Guangdong Pearl Rivers Talent Plan under Grant No. 2019JC01X235, China. We sincerely thank all the editors and anonymous reviewers for their careful work and thoughtful suggestions that have helped improve this paper substantially.

Data availability

Data will be made available on request.

References

- [1] H. Yu, X. Cheng, K.H.M. Cheng, W. Peng, Z. Yu, G. Zhao, Discovering attention-guided cross-modality correlation for visible-infrared person re-identification, *Pattern Recognit.* 155 (2024) 110643, <http://dx.doi.org/10.1016/j.patcog.2024.110643>.
- [2] J. Wu, J. Liu, Y. Liu, L. Wang, Z. Wang, W. Chen, Z. Tian, R. Yu, V.C.M. Leung, A survey on cloud-edge-terminal collaborative intelligence in AIoT networks, 2025, <http://dx.doi.org/10.48550/arXiv.2508.18803>, arXiv:2508.18803.
- [3] W. Ding, T. Hou, J. Huang, H. Ju, S. Jiang, Dynamic evidence fusion neural networks with uncertainty theory and its application in brain network analysis, *Inf. Sci.* 691 (2025) 121622.
- [4] Y. Feng, F. Chen, J. Yu, Y. Ji, F. Wu, S. Liu, X.-Y. Jing, Homogeneous and heterogeneous relational graph for visible-infrared person re-identification, *Pattern Recognit.* 158 (2025) 110981, <http://dx.doi.org/10.1016/j.patcog.2024.110981>.
- [5] X. Zhai, X. Wang, B. Mustafa, A. Steiner, D. Keyzers, A. Kolesnikov, L. Beyer, Lit: Zero-shot transfer with locked-image text tuning, in: *CVPR*, 2022, pp. 18123–18133.
- [6] T. Gong, J. Wang, L. Zhang, Cross-modal semantic aligning and neighbor-aware completing for robust text–image person retrieval, *Inf. Fusion* 112 (2024) 102544.
- [7] J. Liu, Y. Du, K. Yang, J. Wu, Y. Wang, X. Hu, Z. Wang, Y. Liu, P. Sun, A. Boukerche, V.C.M. Leung, Edge-cloud collaborative computing on distributed intelligence and model optimization: A survey, *IEEE Commun. Surv. Tuts.* 28 (2026) 5049–5080, <http://dx.doi.org/10.1109/COMST.2026.3669216>.
- [8] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, X. Wang, Person search with natural language description, in: *CVPR*, 2017, pp. 1970–1979.
- [9] J. Liu, Y. Liu, J. Lin, J. Li, L. Cao, P. Sun, B. Hu, L. Song, A. Boukerche, V.C. Leung, Networking systems for video anomaly detection: A tutorial and survey, *ACM Comput. Surv.* 57 (10) (2025) 270:1–270:37, <http://dx.doi.org/10.1145/3729222>.
- [10] C. Yu, X. Liu, J. Zhu, Y. Wang, P. Zhang, H. Lu, CLIMB-ReID: A hybrid CLIP-Mamba framework for person re-identification, *AAAI* 39 (9) (2025) 9589–9597, <http://dx.doi.org/10.1609/aaai.v39i9.33039>.
- [11] J. Ren, K. Hua, Y. Cao, Global optimal K-medoids clustering of one million samples, in: *NeurIPS*, Vol. 35, 2022, pp. 982–994.
- [12] Y. Qian, Y. Wang, J. Liu, Q. Zou, Y. Ding, X. Guo, W. Ding, A survey on multi-view fusion for predicting links in biomedical bipartite networks: Methods and applications, *Inf. Fusion* 117 (2025) 102894.
- [13] Y. Bai, J. Wang, M. Cao, C. Chen, Z. Cao, L. Nie, M. Zhang, Text-based person search without parallel image-text data, in: *ACM MM*, 2023, pp. 757–767.
- [14] Z. Wang, A. Zhu, J. Xue, X. Wan, C. Liu, T. Wang, Y. Li, CAIBC: Capturing all-round information beyond color for text-based person retrieval, in: *ACM MM*, 2022, pp. 5314–5322.
- [15] J. Ren, V. Osuna-Enciso, M. Okamoto, Q. Mao, C. Ji, L. Cao, K. Hua, Y. Cao, A GPU-accelerated moving-horizon algorithm for training deep classification trees on large datasets, 2023, <http://dx.doi.org/10.48550/arXiv.2311.06952>, arXiv:2311.06952.
- [16] L. Cao, J. Wang, J. Su, Y. Luo, Y. Cao, R.D. Braatz, B. Gopaluni, Comprehensive analysis on machine learning approaches for interpretable and stable soft sensors, *IEEE Trans. Instrum. Meas.* 74 (2025) 1–17, <http://dx.doi.org/10.1109/TIM.2025.3556830>.
- [17] A. Zhu, Z. Wang, Y. Li, X. Wan, J. Jin, T. Wang, F. Hu, G. Hua, DSSL: Deep surroundings-person separation learning for text-based person retrieval, in: *ACM MM*, 2021, pp. 209–217.
- [18] C. Gao, G. Cai, X. Jiang, F. Zheng, J. Zhang, Y. Gong, P. Peng, X. Guo, X. Sun, Contextual non-local alignment over full-scale representation for text-based person search, 2021, arXiv preprint [arXiv:2101.03036](https://arxiv.org/abs/2101.03036).
- [19] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, Y.-D. Shen, Dual-path convolutional image-text embeddings with instance loss, *ACM Trans. Multimed. Comput. Commun. Appl.* 16 (2) (2020) 1–23.
- [20] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in: *ECCV*, 2018, pp. 686–701.
- [21] J. Liu, Z. Ma, H. Yu, B. Ju, W. Yang, C. Li, B. Hu, L. Song, Collaborative adaptive curriculum for progressive knowledge distillation, in: *IEEE ICME*, 2026, <http://dx.doi.org/10.48550/arXiv.2603.20296>.
- [22] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, S.C.H. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, in: *NeurIPS*, Vol. 34, 2021, pp. 9694–9705.
- [23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: *NeurIPS*, Vol. 30, 2017.
- [24] J. Liu, Z. Guo, Y. Wang, X. Zhu, Y. Du, Z. Wang, V.C.M. Leung, Diffusion-guided semantic consistency for multimodal heterogeneity, in: *IEEE ICME*, 2026, <http://dx.doi.org/10.48550/arXiv.2603.19337>.
- [25] Z. Wang, Z. Fang, J. Wang, Y. Yang, Vitaa: Visual-textual attributes alignment in person search by natural language, in: *ECCV*, Springer, 2020, pp. 402–420.
- [26] Z. Ding, C. Ding, Z. Shao, D. Tao, Semantically self-aligned network for text-to-image part-aware person re-identification, 2021, arXiv preprint [arXiv:2107.12666](https://arxiv.org/abs/2107.12666).
- [27] Y. Chen, G. Zhang, Y. Lu, Z. Wang, Y. Zheng, TIPCB: A simple but effective part-based convolutional baseline for text-based person search, *Neurocomputing* 494 (2022) 171–181.
- [28] S. Li, M. Cao, M. Zhang, Learning semantic-aligned feature representation for text-based person search, in: *ICASSP*, IEEE, 2022, pp. 2724–2728.

- [29] C. Wang, Z. Luo, Y. Lin, S. Li, Text-based person search via multi-granularity embedding learning, in: *IJCAI*, 2021, pp. 1068–1074.
- [30] X. Shu, W. Wen, H. Wu, K. Chen, Y. Song, R. Qiao, B. Ren, X. Wang, See finer, see more: Implicit modality alignment for text-based person retrieval, in: *ECCV Wkshps*, Springer, 2023, pp. 624–641.
- [31] S. Yan, N. Dong, L. Zhang, J. Tang, CLIP-driven fine-grained text-image person re-identification, *IEEE Trans. Image Process.* 32 (2023) 6032–6046.
- [32] Z. Ji, J. Hu, D. Liu, L.Y. Wu, Y. Zhao, Asymmetric cross-scale alignment for text-based person search, *IEEE Trans. Multimed.* (2022).
- [33] J. Liu, H. Yu, G. Fang, J. Wu, L. Gong, X. Zhu, Y. Liu, P. Sun, Semantic prototype learning for cross-domain activity recognition, in: *IEEE ICME*, 2026.
- [34] J. Liu, L. Gong, J. Guo, J. Wu, L. Sun, Y. Bi, K. Patwari, B. Chen, L. Zhang, W. Zhou, Y. Liu, X. Zhu, C.-N. Chuah, B. Rajaratnam, Multimodal large language models in medicine and nursing: A survey, 2025, <http://dx.doi.org/10.36227/techrxiv.175623882.21520632/v1>.
- [35] D. Liu, L. Wu, F. Zheng, L. Liu, M. Wang, Verbal-person nets: Pose-guided multi-granularity language-to-person generation, *IEEE Trans. Neural Netw. Learn. Syst.* (2022).
- [36] S. Xie, C. Zhang, R. Ma, Z. Li, Z. Wang, C. Wei, Multi-scale feature refinement via perspective scaling and adaptive regularization for text-based person search, *Eng. Appl. Artif. Intell.* 159 (2025) 111496, <http://dx.doi.org/10.1016/j.engappai.2025.111496>.
- [37] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: *ICML, PMLR*, 2021, pp. 8748–8763.
- [38] Z. Wang, A. Zhu, J. Xue, D. Jiang, C. Liu, Y. Li, F. Hu, SUM: Serialized updating and matching for text-based person retrieval, *Knowl.-Based Syst.* 248 (2022) 108891.
- [39] S. Yan, H. Tang, L. Zhang, J. Tang, Image-specific information suppression and implicit local alignment for text-based person search, *IEEE Trans. Neural Netw. Learn. Syst.* 35 (12) (2024) 17973–17986.
- [40] A. Farooq, M. Awais, J. Kittler, S.S. Khalid, AXM-Net: cross-modal context sharing attention network for person Re-ID, 2021, arXiv preprint [arXiv:2101.08238](https://arxiv.org/abs/2101.08238).
- [41] X. Han, S. He, L. Zhang, T. Xiang, Text-based person search with limited data, 2021, arXiv preprint [arXiv:2110.10807](https://arxiv.org/abs/2110.10807).
- [42] G. Zhang, Y. Chen, Y. Zheng, G. Martin, R. Wang, Local-enhanced representation for text-based person search, *Pattern Recognit.* 161 (2025) 111247, <http://dx.doi.org/10.1016/j.patcog.2024.111247>.
- [43] K. Niu, L. Huang, Y. Long, Y. Huang, L. Wang, Y. Zhang, Comprehensive attribute prediction learning for person search by language, *IEEE Trans. Image Process.* 33 (2024) 1990–2003, <http://dx.doi.org/10.1109/TIP.2024.3372832>.
- [44] D. Jiang, M. Ye, Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval, in: *CVPR*, 2023, pp. 2787–2797.
- [45] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: *CVPR*, 2015, pp. 815–823.
- [46] Y. Sun, C. Cheng, Y. Zhang, C. Zhang, L. Zheng, Z. Wang, Y. Wei, Circle loss: A unified perspective of pair similarity optimization, in: *CVPR*, 2020, pp. 6398–6407.
- [47] J. Ren, N. You, K. Hua, C. Ji, Y. Cao, A global optimization algorithm for K-center clustering of one billion samples, *Manag. Sci.* 72 (6) (2026) 5106–5130, <http://dx.doi.org/10.1287/mnsc.2023.00218>.
- [48] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805).
- [49] Y. Lee, J. Kim, P. Kang, LAnoBERT: System log anomaly detection based on BERT masked language model, *Appl. Soft Comput.* 146 (2023) 110689.
- [50] L. Cao, J. Su, E. Conde, L.C. Siang, Y. Cao, B. Gopaluni, A novel automated soft sensor design tool for industrial applications based on machine learning, *Control Eng. Pract.* 160 (2025) 106322, <http://dx.doi.org/10.1016/j.conengprac.2025.106322>.
- [51] Y. Liu, J. Liu, C. Li, R. Xi, W. Li, L. Cao, J. Wang, L.T. Yang, J. Yuan, W. Zhou, Anomaly detection and generation with diffusion models: A survey, 2025, <http://dx.doi.org/10.48550/arXiv.2506.09368>, [arXiv:2506.09368](https://arxiv.org/abs/2506.09368).
- [52] J. Liu, Z. Ma, Z. Wang, Y. Liu, Z. Wang, P. Sun, L. Song, B. Hu, V.C.M. Leung, A survey on diffusion models for anomaly detection, 2025, <http://dx.doi.org/10.48550/arXiv.2501.11430>, [arXiv:2501.11430](https://arxiv.org/abs/2501.11430).
- [53] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929).
- [54] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: *CVPR, IEEE*, 2009, pp. 248–255.
- [55] J. Liu, W. Zhu, D. Li, X. Hu, L. Song, Domain generalization with semi-supervised learning for people-centric activity recognition, *Sci. China Inf. Sci.* 68 (1) (2025) 112103, <http://dx.doi.org/10.1007/s11432-022-3860-y>.