Learning Disentangled Concept Representations for Fine-Grained Text-to-Image Person Re-Identification

Giyeol Kim[1] Jianing Li [2] Xiaobin Liu [3] Chanho Eom[1,*]
[1]*Graduate School of Advanced Imaging Science, Multimedia and Film (GSAIM), Chung-Ang University, Seoul, Republic of Korea*
[2]*Department of Computing, Hong Kong Polytechnic University, Hong Kong, China*
[3]*College of Artificial Intelligence, Nankai University, Tianjin, China*
{giyeolkim, cheom}@cau.ac.kr, tensor.li@polyu.edu.hk, liuxb@nankai.edu.cn

***Abstract*—Text-to-image person re-identification (TIReID) aims to retrieve person images from a large gallery based on natural language descriptions. This task is challenging due to the large modality gap between visual appearances and textual expressions, and the need to capture fine-grained correspondences that distinguish individuals with similar attributes such as color, texture, and clothing style. To address these challenges, we propose DiCo (Disentangled Concept Representation), a framework that performs hierarchical and disentangled cross-modal alignment. DiCo employs a shared slot-based representation, where each slot serves as a part-level anchor across modalities and is further divided into concept blocks representing complementary attributes (e.g., color, texture, shape). This design enables consistent part-level alignment while disentangling semantic concepts across modalities. Experiments on CUHK-PEDES, ICFG-PEDES, and RSTPReid demonstrate that DiCo achieves competitive performance with state-of-the-art methods and provides improved interpretability through explicit slot- and concept-level representations for fine-grained retrieval.**

***Index Terms*—Text-to-image person re-identification, Slot attention**

I. Introduction

Text-to-image person re-identification (TIReID) aims to retrieve a particular individual within a large-scale image gallery using only a natural language description [1]–[6]. Unlike conventional image-based ReID methods [7]–[13], TIReID allows user queries to be formulated directly in natural language, thereby providing a more flexible and user-friendly retrieval paradigm. Despite its practicality, TIReID remains a highly challenging task due to two fundamental difficulties. (1) A substantial modality gap exists between textual and visual representations, stemming from their inherently heterogeneous nature. Visual data comprise dense, continuous signals that encode variations in pose, viewpoint, and illumination, whereas textual inputs consist of discrete linguistic tokens expressing abstract semantic concepts. This disparity makes direct alignment across the two modalities non-trivial, underscoring the importance of learning a shared embedding space for robust cross-modal matching. (2) Another key challenge involves capturing fine-grained correspondences between textual attributes and subtle visual cues. Because part-level annotations are rarely available in real-world datasets, it is crucial to design models capable of automatically discovering and aligning discriminative local features without explicit supervision.

To tackle the aforementioned challenges, early approaches [1]–[4], [14], [15] commonly employed dual-encoder frameworks, where both images and texts are mapped into a unified embedding space through global feature representations and metric learning objectives. However, this global alignment strategy often fails to capture fine-grained and discriminative local details. To address this limitation, subsequent part-based or region-level methods [5], [6], [16]–[20] decomposed person images into body parts and aligned these regions with word-level textual tokens, thereby establishing more localized and semantically consistent correspondences between the two modalities. More recently, vision-language models (VLMs) [21] have been adopted to enhance textual comprehension and contextual reasoning, leading to richer semantic embeddings that significantly improve retrieval performance [18], [22], [23]. Despite these advances, existing methods still struggle to fundamentally reconcile the heterogeneous characteristics of visual and textual modalities. To mitigate this issue, prior works [1]–[6], [14]–[18], [22], [23] have sought to reduce the modality gap either by enforcing metric-based alignment between dual-encoder features in a shared embedding space or by incorporating cross-attention mechanisms atop dual-encoder representations. Although such techniques enable limited cross-modal interactions, they tend to yield shallow feature fusion rather than deep semantic alignment. Moreover, most existing alignment strategies [5], [6], [16]–[19] remain constrained to global identity features or coarse part-level correspondences, making it difficult to explicitly disentangle nuanced semantic factors such as color, texture, and shape. As a result, subtle yet discriminative cues that are crucial for distinguishing visually similar individuals are often neglected.

In this work, we introduce DiCo (Disentangled Concept Representation), a novel framework designed to establish hierarchical and interpretable cross-modal alignments through structured slot decomposition and concept-level disentangle-

*Corresponding author.

ment. DiCo employs a unified collection of learnable slots that serve as modality-shared anchors corresponding to distinct body regions. Each slot is further decomposed into several concept blocks that encapsulate semantically coherent attributes such as color, texture, and shape. Unlike prior methods relying on shallow feature fusion, DiCo bridges the intrinsic modality gap through iterative and attentive interactions that progressively refine semantic consistency between visual and textual representations. By hierarchically structuring the representation space—from global identity cues down to concept-specific subspaces—DiCo can automatically discover and align subtle visual attributes with their textual counterparts, even in the absence of explicit part-level supervision. This hierarchical design enables the model to reason effectively across both coarse and fine-grained semantics, thereby improving its ability to discriminate between individuals with subtle attribute variations. Furthermore, DiCo is trained under multi-level contrastive objectives and reconstruction constraints, ensuring alignment at global, part, and concept levels. Through this multi-granularity supervision, DiCo learns representations that are both robust and semantically grounded. Comprehensive experiments conducted on three public benchmarks demonstrate the superiority of our method, showing competitive or better performance compared to state-of-the-art approaches.

II. Related Works

A. Text-to-image Person Re-identification

Text-to-image person re-identification (TIReID) aims to retrieve the correct person image from a large gallery based on a free-form textual description. Early methods [1]–[4], [14], [15] typically adopt dual-encoder architectures, where images and texts are embedded into a shared space using global alignment objectives. While such methods establish a foundation for cross-modal retrieval, they struggle to capture fine-grained attributes essential for distinguishing visually similar individuals. To address this issue, subsequent methods [5], [6], [16]–[20] introduce part-based models that decompose person images into body regions and align them with word-level tokens, enabling more localized correspondences. Despite these advances, existing methods still face fundamental limitations in modeling fine-grained cross-modal correspondences. While these methods [5], [6], [16]–[20] have enabled more localized alignment, the extracted part features are typically entangled representations that fail to separate key semantic attributes like color, texture, and shape within each region. Such entangled representations hinder precise alignment, as they obscure subtle yet discriminative cues necessary for identifying visually similar individuals.

B. Slot Attention

Slot Attention [24] is proposed for unsupervised object-centric learning, aiming to decompose visual scenes into slot-based latent representations through an iterative attention mechanism. Each slot typically corresponds to an object or a part, yielding structured and interpretable representations without requiring explicit supervision. Subsequent methods extend Slot Attention to capture temporal dynamics in videos [25]–[28], or integrate top-down signals to improve stability and semantic alignment [29]. While slots provide region-level representations, the semantic concepts within each region remain entangled, which limits their ability to capture fine-grained correspondences. In contrast, our DiCo extends the slot-based paradigm by introducing modality-shared slots that serve as common anchors across vision and language, and further factorizing each slot into disentangled concept representations.

III. Proposed Method

A. Overall Framework

Given a person image I and a free-form textual description T, the goal of TIReID is to retrieve person images matching a given textual description by learning discriminative and semantically aligned cross-modal representations. For the visual branch, a visual backbone processes the image into a set of patch tokens $\mathbf{X}_v = \{x_n^v\}_{n=1}^N \in \mathbb{R}^{N\times d}$ and a global representation $g_v \in \mathbb{R}^d$, where N denotes the number of image patches and d is the feature embedding dimension. For the textual branch, a textual backbone encodes the description into word or subword tokens $\mathbf{X}_t = \{x_\ell^t\}_{\ell=1}^L \in \mathbb{R}^{L\times d}$ and a global representation $g_t \in \mathbb{R}^d$, where L is the number of tokens in the text sequence.

To achieve fine-grained cross-modal alignment, we introduce a shared set of K learnable slots $\mathbf{S}^{(0)} = \{s_k^{(0)}\}_{k=1}^K$, where each slot is initialized as $s_k^{(0)} \in \mathbb{R}^{M\times d_c}$. Each slot is further factorized into M concept blocks such that $s_k = [\, s_{k,1}; \ldots; s_{k,M} \,]$ with $s_{k,m} \in \mathbb{R}^{d_c}$. The same slot indices are shared across the visual and textual branches, encouraging the k-th slot to capture a consistent semantic body part in the person image and the corresponding descriptive expression in the text. Meanwhile, the concept blocks within each slot provide a disentangled representation of attributes such as color, texture, and shape. Starting from $\mathbf{S}^{(0)}$, each modality refines its slots through an iterative interaction module, which aggregates tokens into slots and updates their representations. After T refinement steps, we obtain the final slot embeddings:

$$\mathbf{S}_v^{(T)} = \{s_k^{v,(T)}\}_{k=1}^K, \qquad \mathbf{S}_t^{(T)} = \{s_k^{t,(T)}\}_{k=1}^K, \tag{1}$$

where $s_k^{v,(T)}$ and $s_k^{t,(T)}$ denote the k-th refined slot for the visual and textual branches.

The model outputs both global embeddings $\big(g_v, g_t\big)$ and refined slot embeddings $\big(\mathbf{S}_v^{(T)}, \mathbf{S}_t^{(T)}\big)$, which jointly contribute complementary information for TIReID. During training, the model is optimized with multi-level objectives that encourage global alignment between global representations, part-level alignment across slots, and concept-level alignment within disentangled blocks. During inference, we compute similarities by jointly comparing the three levels of representation: global embeddings for identity cues, slot embeddings for region-level alignment, and block embeddings for fine-grained attribute matching.

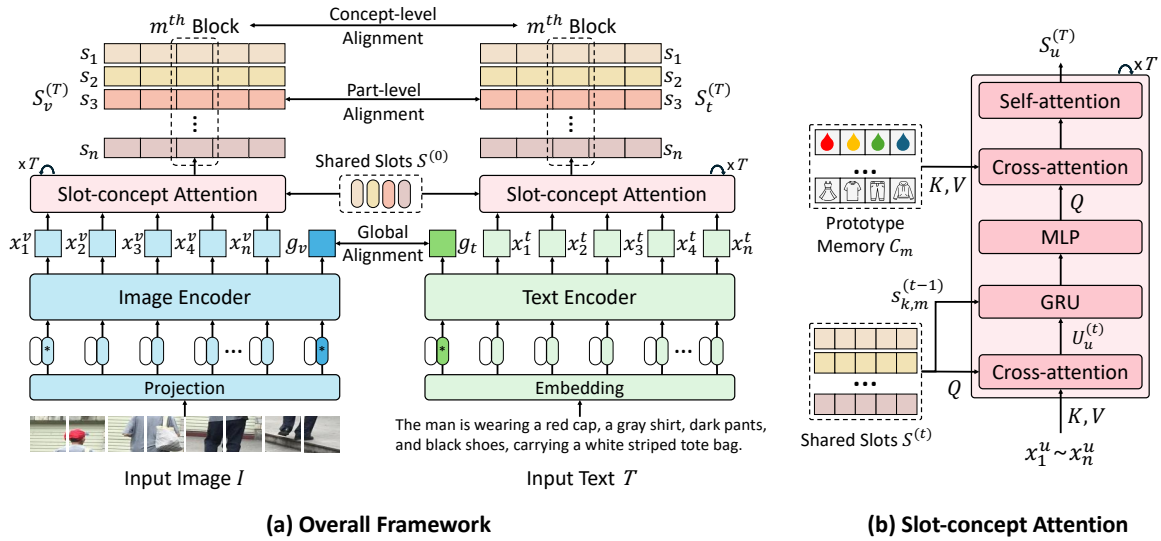

Fig. 1. **(a) Overall framework:** Given an input image and text, visual and textual features are extracted by respective encoders, followed by global alignment and refinement through shared slots. The refined slot representations are aligned at both part- and concept-levels to capture fine-grained cross-modal correspondences. **(b) Slot-concept Attention:** disentangles slot representations into interpretable concept blocks while ensuring semantic consistency across modalities.

B. Disentangled Concept Representation

To achieve fine-grained cross-modal alignment, we employ a shared set of K learnable slots, where each slot $s_k \in \mathbb{R}^{M \times d_c}$ is decomposed into M concept blocks $s_{k,m} \in \mathbb{R}^{d_c}$. Slots act as part-level anchors across modalities, while their blocks disentangle complementary attributes such as color, texture, or shape. Starting from the initialization $\mathbf{S}^{(0)} = \{s_k^{(0)}\}_{k=1}^{K}$, tokens from both the visual and textual branches are iteratively aggregated into slots and refined block by block, encouraging consistent semantics without explicit supervision.

At refinement step t, input tokens from each modality are assigned to slots through a cross-attention mechanism. Specifically, for modality $u \in v, t$, we obtain queries from the slots $q_u(\mathbf{S}^{(t-1)})$ and keys/values from the input tokens $k_u(\mathbf{X}_u), v_u(\mathbf{X}_u)$, where q_u, k_u, and v_u denote modality-specific learnable linear projections. The attention score matrix is then computed as

$$\mathbf{M}_u^{(t)} = \frac{k_u(\mathbf{X}_u)\, q_u(\mathbf{S}^{(t-1)})^\top}{\sqrt{d_h}} \in \mathbb{R}^{|\mathbf{x}_u| \times K}, \tag{2}$$

which represents the similarity between each input token and each slot. We apply a softmax over slots to assign tokens competitively, followed by a normalization across tokens to ensure balanced slot coverage:

$$\mathbf{A}_u^{(t)} = \mathrm{softmax}_k(\mathbf{M}_u^{(t)}), \qquad \tilde{\mathbf{A}}_u^{(t)} = \mathbf{A}_u^{(t)} / (\mathbf{A}_u^{(t)} \mathbf{1}). \tag{3}$$

The aggregated token readouts are then obtained as

$$\mathbf{U}_u^{(t)} = \left(\tilde{\mathbf{A}}_u^{(t)}\right)^\top v_u(\mathbf{X}_u) \in \mathbb{R}^{K \times (M \times d_c)}, \tag{4}$$

where each row $u_k^{(t)}$ represents the readout of slot k. The readout serves as a slot-specific summary vector obtained by aggregating the input tokens assigned to the slot, thereby capturing the localized information that the slot has attended to in modality u. Each slot readout $u_k^{(t)}$ is partitioned into $\{u_{k,m}^{(t)}\}_{m=1}^{M}$, corresponding to its concept blocks. This decomposition allows each block to focus on a distinct semantic factor (*e.g.*, color, texture, or shape), rather than keeping all attributes entangled within a single vector. Each block is refined independently through a gated residual update:

$$\begin{aligned} \hat{s}_{k,m}^{(t)} &= \mathrm{GRU}_m(s_{k,m}^{(t-1)}, u_{k,m}^{(t)}), \\ \bar{s}_{k,m}^{(t)} &= \hat{s}_{k,m}^{(t)} + \mathrm{MLP}_m(\mathrm{LN}(\hat{s}_{k,m}^{(t)})). \end{aligned} \tag{5}$$

By refining blocks independently, the model achieves a more structured representation where different modules specialize in complementary aspects of appearance. To encourage separation of semantic factors, each block is projected onto a shared prototype memory $\mathbf{C}_m \in \mathbb{R}^{K_m \times d_c}$:

$$s_{k,m}^{(t)} = \mathrm{softmax}\left(\frac{\bar{s}_{k,m}^{(t)} \mathbf{C}_m^\top}{\sqrt{d_c}}\right) \mathbf{C}_m. \tag{6}$$

$\mathbf{C}_m$ serves as a compact dictionary of semantic prototypes that are shared across modalities. Each block maintains its own independent prototype memory $\mathbf{C}_m$, ensuring that different semantic factors are represented in distinct subspaces. By mapping block features to a small set of prototypes, the model enforces each block to be represented in a discrete and interpretable semantic subspace. This prototype projection not only regularizes the learning of disentangled concept factors but also ensures that blocks with the same index in image and text are aligned to consistent semantic anchors, thereby facilitating fine-grained cross-modal correspondence.

After block refinement, each slot is reassembled as $s_k^{(t)} = [\, s_{k,1}^{(t)}; \ldots; s_{k,M}^{(t)} \,]$. We then apply a lightweight self-attention across slots to reduce redundancy among adjacent regions while preserving their permutation consistency. Importantly, both the initial slots $\mathbf{S}^{(0)}$ and the concept memories $\{\mathbf{C}_m\}_{m=1}^{M}$ are shared across modalities, encouraging the k-th slot to

represent the same semantic region in images and the corresponding description in text. After T refinement steps, we obtain the final modality-specific slots:

$$\mathbf{S}_v^{(T)} = \{s_k^{v,(T)}\}_{k=1}^K, \qquad \mathbf{S}_t^{(T)} = \{s_k^{t,(T)}\}_{k=1}^K, \tag{7}$$

where each slot is further decomposed into block embeddings $s_{k,m}^{u,(T)} \in \mathbb{R}^{d_c}$. These structured outputs provide a unified representation that captures part-level anchors and disentangled concept-level semantics, enabling fine-grained and interpretable cross-modal correspondence.

C. Training and Inference

The training objective aims to align visual and textual representations at different levels of granularity. At the global level, we optimize a bidirectional contrastive loss between the image embedding g_v and the text embedding g_t to capture overall cross-modal correspondence. This pulls paired samples closer in the embedding space while pushing non-matching pairs apart:

$$\begin{aligned}\mathcal{L}_{\text{global}} = -\frac{1}{B}\sum_{i=1}^{B}\Bigg[&\log \frac{\exp(\langle g_v^i, g_t^i\rangle/\tau)}{\sum_{j=1}^{B}\exp(\langle g_v^i, g_t^j\rangle/\tau)} \\ &+ \log \frac{\exp(\langle g_t^i, g_v^i\rangle/\tau)}{\sum_{j=1}^{B}\exp(\langle g_t^i, g_v^j\rangle/\tau)}\Bigg],\end{aligned} \tag{8}$$

where τ is a learnable temperature.

At the slot level, we align the refined slots across modalities, ensuring that the k-th slot in the visual branch corresponds to the same semantic region as the k-th slot in the textual branch:

$$\mathcal{L}_{\text{slot}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K}\left[\log \frac{\exp(\langle z_k^{v,i}, z_k^{t,i}\rangle/\tau_s)}{\sum_{j=1}^{B}\exp(\langle z_k^{v,i}, z_k^{t,j}\rangle/\tau_s)}\right]. \tag{9}$$

At the block level, we encourage alignment of disentangled concepts. Since each slot is further decomposed into concept blocks (*e.g.*, color, texture, shape), we align corresponding blocks across modalities to ensure consistent matching of fine-grained cues:

$$\mathcal{L}_{\text{block}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K}\sum_{m=1}^{M}\log \frac{\exp(\langle z_{k,m}^{v,i}, z_{k,m}^{t,i}\rangle/\tau_b)}{\sum_{j=1}^{B}\exp(\langle z_{k,m}^{v,i}, z_{k,m}^{t,j}\rangle/\tau_b)}. \tag{10}$$

To further enhance discriminability, we introduce identity supervision at both global and local levels. At the global level, the embeddings g_v and g_t are classified into identity labels, ensuring that they remain discriminative beyond cross-modal matching:

$$\mathcal{L}_{\text{ID}}^{\text{global}} = -\frac{1}{B}\sum_{i=1}^{B}\big[\log P(y_i \mid g_v^i) + \log P(y_i \mid g_t^i)\big]. \tag{11}$$

At the local level, each slot embedding z_k^u ($u \in v, t$) is supervised with the same identity label, encouraging part-level features to preserve identity cues:

$$\mathcal{L}_{\text{ID}}^{\text{slot}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K}\big[\log P(y_i \mid z_k^{v,i}) + \log P(y_i \mid z_k^{t,i})\big]. \tag{12}$$

Finally, we introduce a reconstruction objective to stabilize slot refinement. By reconstructing the original token features from aggregated slots, the model is encouraged to preserve the fine-grained information captured in the inputs:

$$\mathcal{L}_{\text{rec}} = \frac{1}{B}\sum_{i=1}^{B}\big(\|\hat{\mathbf{X}}_v^i - \mathbf{X}_v^i\|_2^2 + \|\hat{\mathbf{X}}_t^i - \mathbf{X}_t^i\|_2^2\big). \tag{13}$$

For clarity, we group these objectives into three components: alignment loss $\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{global}} + \lambda_s\mathcal{L}_{\text{slot}} + \lambda_b\mathcal{L}_{\text{block}}$, identity loss $\mathcal{L}_{\text{ID}} = \mathcal{L}_{\text{ID}}^{\text{global}} + \mathcal{L}_{\text{ID}}^{\text{local}}$, and reconstruction loss $\mathcal{L}_{\text{rec}}$. The overall training objective is then expressed as

$$\mathcal{L} = \mathcal{L}_{\text{align}} + \mathcal{L}_{\text{ID}} + \lambda_r\mathcal{L}_{\text{rec}}. \tag{14}$$

During inference, a text query is matched with gallery images by combining similarities across multiple levels. Global embeddings provide identity-level cues, refined slots capture region-level correspondences, and block embeddings align fine-grained semantic attributes. The final retrieval score is computed as a weighted combination of these similarities, balancing overall identity with disentangled concept-level details.

IV. Experiments

A. Experiment Setting

1) Datasets: We conduct experiments on three widely used benchmarks for text-based person retrieval: CUHK-PEDES [30], ICFG-PEDES [31], and RSTPReid [32]. CUHK-PEDES contains 40,206 images of 13,003 identities, each paired with two natural language descriptions, totaling 80,440 sentences. It is split into 34,054 images for training, 3,078 for validation, and 3,074 for testing. ICFG-PEDES is constructed from the ICFG-Person dataset and comprises 54,522 images of 4,102 identities, each annotated with one descriptive sentence. The official split includes 34,674 training images, 19,651 test images, and 197 for validation. RSTPReid is a large-scale benchmark focusing on more challenging scenarios. It consists of 20,505 images of 4,101 identities, with each image annotated by a single textual description.

2) Implementation Details: Our model is built upon the CLIP [21] ViT-L/14 [33] backbone with an image resolution of 384×128 pixels and a stride size of 16. The slot attention mechanism employs 8 slots with a slot dimension of 2,048, organized into 8 blocks of 256 dimensions each. We utilize a shared prototype memory with 512 prototypes across modalities to facilitate cross-modal part correspondence learning. The slot attention module performs 5 iterations during the iterative refinement process, and the temperature parameter for prototype assignment is set to 1.0. For training, we use the Adam [34] optimizer with an initial learning rate of 5e-6. The overall loss is a weighted sum of all objectives, where λ_s and λ_b are set to 0.5 for the slot- and block-level losses, respectively, and λ_b is set to 0.1. At inference, the final score is obtained from a weighted combination of global-, slot-, and block-level similarities, with all weights set to 1.

TABLE I

PERFORMANCE OF TEXT-TO-IMAGE PERSON RE-IDENTIFICATION METHODS ON THREE BENCHMARK DATASETS. **BOLD** DENOTES THE BEST AND UNDERLINE DENOTES THE SECOND BEST.

Method	CUHK-PEDES		ICFG-PEDES		RSTPReid	
	R@1	R@5	R@1	R@5	R@1	R@5
GNA-RNN [14]	19.05	-	-	-	-	-
CMPM/C [4]	49.37	71.69	43.51	65.44	-	-
PMA [35]	53.81	73.54	-	-	-	-
TIMAM [3]	54.51	77.56	-	-	-	-
ViTAA [36]	55.97	75.84	50.98	68.79	-	-
NAFS [16]	59.94	79.86	-	-	-	-
SSAN [6]	61.37	80.15	54.23	72.63	43.50	67.80
SRCF [19]	64.04	82.99	57.18	75.01	-	-
TIPCB [5]	64.26	83.19	-	-	-	-
SAF [37]	64.13	82.62	-	-	-	-
IVT [38]	65.59	83.11	56.04	73.60	46.70	70.00
CFine [23]	69.57	85.93	60.83	75.55	50.55	72.50
IRRA [22]	73.38	89.93	63.46	80.24	60.20	81.30
PLOT [18]	75.28	90.42	65.76	81.39	61.80	82.85
Ours	**75.79**	**91.10**	**66.20**	**82.18**	**62.79**	**83.96**

B. Quantitative Results

We compare our DiCo with state-of-the-art text-to-image person re-identification methods on CUHK-PEDES [30], ICFG-PEDES [31], and RSTPReid [32]. As shown in Table I, methods that perform only global matching (*e.g.*, GNA-RNN [14], CMPM/C [4]) fail to capture discriminative details, leading to inferior performance. Part-level matching methods (*e.g.*, ViTAA [36], IRRA [22], PLOT [18]) improve discriminability by localizing body regions, but they still fail to disentangle concept-specific factors such as color or clothing type, leading to ambiguous alignments and reduced robustness. In contrast, our proposed DiCo performs hierarchical alignment across global, slot, and block levels, explicitly balancing identity cues with fine-grained semantic concepts.

C. Qualitative Results.

a) Slot-Level Attention.: In Fig. 2, we provide visualizations of the learned slot-level attention to better understand how our model captures fine-grained correspondences between text and image. Each slot attends to distinct regions of the person image, such as the shirt, pants, or shoes, while simultaneously focusing on semantically related words in the textual description. For example, slots specialize in attributes like *"yellow shirt"*, *"floral pants"*, or *"blue shirt"*, consistently linking localized visual cues with their linguistic counterparts. These results highlight that the proposed disentangled slot representations partition the image into semantically meaningful parts. They also capture fine-grained alignments with textual tokens, enabling precise cross-modal matching without the need for part-level annotations.

b) Block-Level Clustering.: We validate the disentanglement ability of DiCo by analyzing block embeddings on CUHK-PEDES [30], as shown in Fig.3. Specifically, we randomly sample 10 person images and visualize the embeddings of 8 blocks within the second slot using t-SNE [39]. As shown in Fig. 3, features from different samples cluster together when they share the same concept, even across distinct individuals. This indicates that each block consistently captures the same underlying semantic concept across samples, rather than being entangled with individual appearance differences. This clustering clearly demonstrates that the proposed block-level design learns concept-specific representations, facilitating fine-grained and interpretable cross-modal alignment.

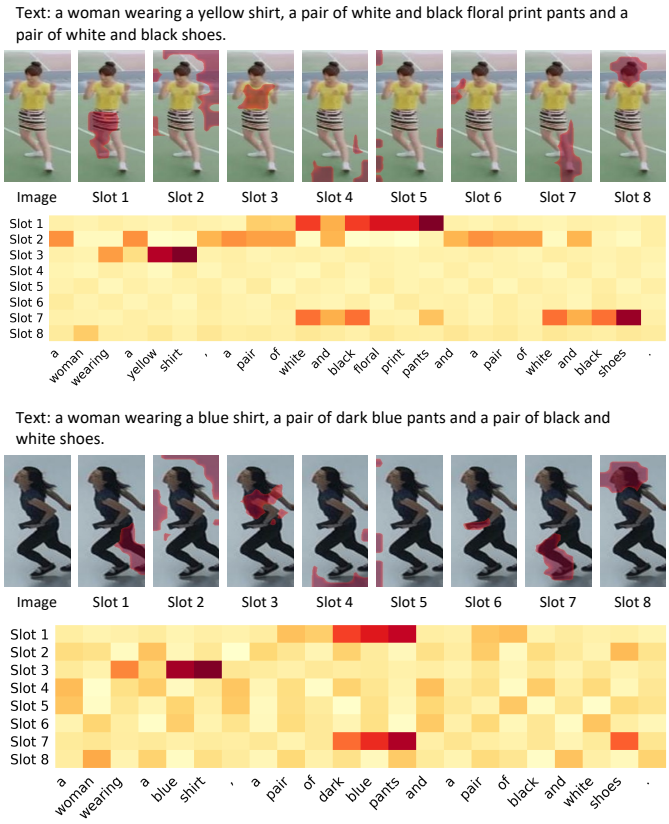

Fig. 2. Visualization of slot-level attention for text queries and corresponding images. Each slot attends to distinct body regions and aligns with relevant words in the query, effectively capturing fine-grained attributes such as color, clothing type, and shoes.

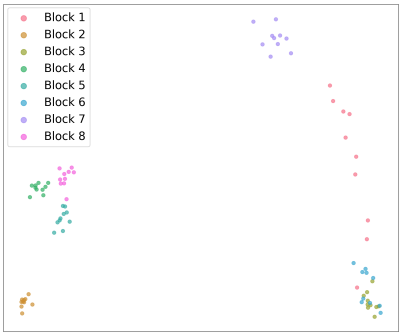

Fig. 3. t-SNE [39] visualization of 8 blocks embeddings from the second slot on randomly sampled images from the CUHK-PEDES [30] test set. Despite originating from different person samples, the embeddings form coherent clusters, indicating that the block consistently captures the same underlying semantic concept across individuals.

ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the Graduate School of Meta-

verse Convergence support program (IITP-2024-RS-2024-00418847) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00355008).

REFERENCES

[1] T. Chen, C. Xu, J. Luo, Improving text-based person search by spatial matching and adaptive threshold, in: Proceedings of the IEEE/CVF Conference on Winter Conference on Applications of Computer Vision, 2018, pp. 1879–1887.

[2] S. Li, T. Xiao, H. Li, W. Yang, X. Wang, Identity-aware textual-visual matching with latent co-attention, in: Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision, 2017, pp. 1890–1899.

[3] N. Sarafianos, X. Xu, I. A. Kakadiaris, Adversarial representation learning for text-to-image matching, in: Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision, 2019, pp. 5814–5824.

[4] Y. Zhang, H. Lu, Deep cross-modal projection learning for image-text matching, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 686–701.

[5] 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) pp. 171–181.

[6] Z. Ding, C. Ding, Z. Shao, D. Tao, Semantically self-aligned network for text-to-image part-aware person re-identification, arXiv preprint arXiv:2107.12666 (2021).

[7] S. He, H. Luo, P. Wang, F. Wang, H. Li, W. Jiang, Transreid: Transformer-based object re-identification, in: Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision, 2021, pp. 15013–15022.

[8] D. Wang, Y. Chen, L. Tao, C. Hu, Z. Tie, W. Ke, Aea-net:affinity-supervised entanglement attentive network for person re-identification, Pattern Recognition Letters 172 (2023) pp. 237–244.

[9] Y. Chen, S. Xia, J. Zhao, Y. Zhou, Q. Niu, R. Yao, D. Zhu, D. Liu, Rest-reid: Transformer block-based residual learning for person re-identification, Pattern Recognition Letters 157 (2022) pp. 90–96.

[10] C. Eom, G. Lee, K. Cho, H. Jung, M. Jin, B. Ham, Cerberus: Attribute-based person re-identification using semantic ids, Expert Systems with Applications 259 (2025) pp. 125320–125336.

[11] H. Lee, J. Park, J. Oh, C. Eom, Domain generalization for person re-identification: A survey towards domain-agnostic person matching, Neurocomputing (2025) pp. 130763–130782.

[12] C. Eom, B. Ham, Learning disentangled representation for robust person re-identification, Advances in Neural Information Processing Systems 32 (2019).

[13] J. Sung, H. Kim, M. Kim, Y. Mok, C. Park, J. Paik, Synthetic image generation for data augmentation to train an unconscious person detection network in a uav environment, IEIE Transactions on Smart Processing & Computing (2022) pp. 156–161.

[14] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, X. Wang, Person search with natural language description, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 1970–1979.

[15] Y. Chen, R. Huang, H. Chang, C. Tan, T. Xue, B. Ma, Cross-modal knowledge adaptation for language-based person search, IEEE Transactions on Image Processing 30 (2021) pp. 4057–4069.

[16] 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, arXiv preprint arXiv:2101.03036 (2021).

[17] K. Niu, Y. Huang, W. Ouyang, L. Wang, Improving description-based person re-identification by multi-granularity image-text alignments, IEEE Transactions on Image Processing 29 (2020) pp. 5542–5556.

[18] J. Park, D. Kim, B. Jeong, S. Kwak, Plot: Text-based person search with part slot attention for corresponding part discovery, in: Proceedings of the European Conference on Computer Vision, 2024, pp. 474–490.

[19] W. Suo, M. Sun, K. Niu, Y. Gao, P. Wang, Y. Zhang, Q. Wu, A simple and robust correlation filtering method for text-based person search, in: Proceedings of the European Conference on Computer Vision, 2022, pp. 726–742.

[20] C. Wang, Z. Luo, Z. Zhong, S. Li, Divide-and-merge the embedding space for cross-modality person search, Neurocomputing 463 (2021) pp. 388–399.

[21] 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: Proceedings of the International Conference on Machine Learning, 2021, pp. 8748–8763.

[22] D. Jiang, M. Ye, 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, 2023, pp. 2787–2797.

[23] S. Yan, N. Dong, L. Zhang, J. Tang, Clip-driven fine-grained text-image person re-identification, IEEE Transactions on Image Processing 32 (2023) pp. 6032–6046.

[24] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, T. Kipf, Object-centric learning with slot attention, Advances in Neural Information Processing Systems 33 (2020) pp. 11525–11538.

[25] T. Kipf, G. F. Elsayed, A. Mahendran, A. Stone, S. Sabour, G. Heigold, R. Jonschkowski, A. Dosovitskiy, K. Greff, Conditional object-centric learning from video, arXiv preprint arXiv:2111.12594 (2021).

[26] K. Fan, Z. Bai, T. Xiao, D. Zietlow, M. Horn, Z. Zhao, C.-J. Simon-Gabriel, M. Z. Shou, F. Locatello, B. Schiele, et al., Unsupervised open-vocabulary object localization in videos, in: Proceedings of the IEEE/CVF Conference on International Conference on Computer Vision, 2023, pp. 13747–13755.

[27] G. Elsayed, A. Mahendran, S. Van Steenkiste, K. Greff, M. C. Mozer, T. Kipf, Savi++: Towards end-to-end object-centric learning from real-world videos, Advances in Neural Information Processing Systems 35 (2022) pp. 28940–28954.

[28] G. Singh, Y.-F. Wu, S. Ahn, Simple unsupervised object-centric learning for complex and naturalistic videos, Advances in Neural Information Processing Systems 35 (2022) pp. 18181–18196.

[29] D. Kim, S. Kim, S. Kwak, Bootstrapping top-down information for self-modulating slot attention, Advances in Neural Information Processing Systems 37 (2024) pp. 103751–103773.

[30] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, X. Wang, Person search with natural language description, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 1970–1979.

[31] Z. Ding, C. Ding, Z. Shao, D. Tao, Semantically self-aligned network for text-to-image part-aware person re-identification, arXiv preprint arXiv:2107.12666 (2021).

[32] 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: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 209–217.

[33] 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, arXiv preprint arXiv:2010.11929 (2020).

[34] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[35] Y. Jing, C. Si, J. Wang, W. Wang, L. Wang, T. Tan, Pose-guided multi-granularity attention network for text-based person search, in: Proceedings of the AAAI conference on Artificial Intelligence, Vol. 34, 2020, pp. 11189–11196.

[36] Z. Wang, Z. Fang, J. Wang, Y. Yang, Vitaa: Visual-textual attributes alignment in person search by natural language, in: Proceedings of the European Conference on Computer Vision, 2020, pp. 402–420.

[37] S. Li, M. Cao, M. Zhang, Learning semantic-aligned feature representation for text-based person search, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2022, pp. 2724–2728.

[38] 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: Proceedings of the European Conference on Computer Vision, 2022, pp. 624–641.

[39] L. van der Maaten, G. Hinton, Visualizing data using t-sne, Journal of Machine Learning Research 9 (86) (2008) pp. 2579–2605.