Compositional Preference Learning for Composed Person Retrieval

Anonymous CVPR submission

Paper ID *****

Abstract

In aerial surveillance, person retrieval is challenging due to low resolution, large viewpoint changes, and appearance ambiguity, which limit appearance-only matching. These conditions motivate multimodal person retrieval settings in which identity cues from a reference image are combined with language-based descriptions of appearance change. We study this problem through composed person retrieval (CPR), where a target image is retrieved given a reference image and a modification text. We propose a simple framework that models compositional preference by constructing mismatched compositions via cross-sample replacement of either images or texts, and training the model to rank the original composition above these variants. This objective encourages the model to remain sensitive to changes in either the modification text or the reference identity, enabling finer-grained compositional reasoning. Our method achieves state-of-the-art performance on the ITCPR benchmark, surpassing the previous best supervised CPR baseline by 2.18% in Rank-1 and 1.80% in mAP. These results demonstrate that explicitly modeling compositional preference is an effective strategy for composed person retrieval and a promising direction for challenging surveillance scenarios, particularly aerial surveillance.

1. Introduction

In aerial and cross-view surveillance scenarios [4, 9, 14, 17], person retrieval is particularly challenging due to low resolution, large viewpoint variations, and appearance ambiguity, which often render appearance-only matching unreliable. These challenges motivate retrieval formulations that integrate visual identity cues with language-based descriptions of appearance changes. In this work, we study this problem in the framework of composed person retrieval (CPR), which aims to retrieve a target person image from a gallery given a multimodal query composed of a reference image and a modification text. Unlike conventional person retrieval, CPR [7, 8, 16] is designed for scenarios where the target person should preserve the core identity of the reference while reflecting a specific appearance change described in natural language, such as changes in clothing, accessories, or other visible attributes. As a result, CPR requires both identity preservation and fine-grained compositional reasoning across vision and language.

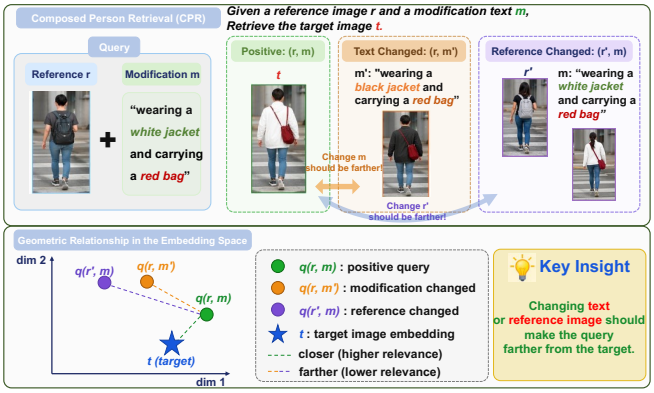

Figure 1. Relational structure in composed person retrieval. The original query should remain more compatible with the target image than mismatched variants constructed by changing either the modification text or the reference image.

The central difficulty of CPR lies in distinguishing *which components* of the multimodal query govern identity preservation and *which part* specifies appearance change, rather than merely aligning the reference image, the text, and the target image at a coarse level. *However, existing CPR methods mainly learn from positive triplets, without explicitly exploiting the richer relational structure among triplets within a mini-batch.*

In this paper, we revisit CPR from the perspective of relational compositional supervision. Our key observation is that samples within a training batch naturally allow the construction of informative mismatched compositions: one can replace the modification text of a triplet with that of another, or replace the reference image while keeping the modification text fixed. Although these variants are semantically close to the original query, they should be less compatible with the corresponding target. This provides a simple yet effective way to introduce finer-grained supervision beyond

standard triplet matching. As illustrated in Fig. 1, changing either the modification text or the reference identity should move the composed query farther away from the correct target in the embedding space.

Based on this intuition, we propose a straightforward framework CPR framework: **CO**mpositional **P**reference learning for compos**E**d person **R**etrieval (COPER). COPER strengthens model training by mining mismatched compositions within each mini-batch and introduces a Compositional Preference Loss that encourages the original query–target pair to be ranked above its corrupted counterparts. By doing so, the model is trained not only to match valid compositions, but also to distinguish them from subtly mismatched alternatives that differ in either appearance modification or identity cue.

Our main contributions are summarized as follows:

- We identify the lack of fine-grained compositional supervision as a key limitation in existing CPR methods, and revisit the task from a relational reasoning perspective within a mini-batch.
- We propose a Compositional Preference Loss based on mismatched composition mining, which constructs mismatched triplets by swapping the reference image or modification text across samples and encourages the model to rank the original composition higher.
- We validate the effectiveness of our method on the ITCPR benchmark, achieving state-of-the-art performance among supervised CPR approaches.

2. Related Work

2.1. Composed Image Retrieval

Composed image retrieval (CIR) extends standard retrieval by using a multimodal query composed of a reference image and a modification text. Existing CIR methods can be divided into supervised and zero-shot approaches. Supervised methods learn from triplet annotations, while zero-shot methods avoid such supervision by leveraging pretrained vision-language models and pseudo-word based composition. Representative zero-shot CIR methods include Pic2Word [10], SEARLE [1], and LinCIR [2]. However, these methods are mainly designed for natural images and do not explicitly address the fine-grained identity sensitivity required in person retrieval.

2.2. Composed Person Retrieval

Composed person retrieval (CPR) extends composed image retrieval to person search, where the goal is to retrieve images of the same person after appearance changes described by text. Word4Per [8] is the first work to formulate CPR and introduce the ITCPR benchmark. I2ID [16] further improves this line by suppressing non-identity appearance cues that may conflict with the modification text. FAFA [7] advances CPR by introducing SynCPR, a large-scale synthetic training dataset, together with an adaptive feature alignment framework.

3. Method

3.1. Task Formulation

Given a dataset $\mathcal{D} = \{(r_i, m_i, t_i)\}_{i=1}^{M}$, where r_i denotes a reference image, m_i a modification text, and t_i the corresponding target image, the goal is to learn a retrieval model that composes the identity-related visual information from r_i with the semantic modification described by m_i. For a mini-batch of size N sampled from $\mathcal{D}$, we denote the sets of reference images, modification texts, and target images by $\mathcal{R} = \{r_i\}_{i=1}^{N}$, $\mathcal{M} = \{m_i\}_{i=1}^{N}$, and $\mathcal{T} = \{t_i\}_{i=1}^{N}$, respectively. Given a query pair (r_i, m_i), the model aims to retrieve the corresponding target image t_i from the candidate set $\mathcal{T}$ such that the retrieved image reflects the modification in m_i while preserving the identity-related characteristics of r_i.

3.2. Overall framework

Our COPER framework is built on the Q-Former architecture from BLIP-2 [5], which enables efficient multimodal interaction between image and text representations while requiring only lightweight trainable modules. Let $E_{\text{img}}(\cdot)$ denote the image encoder and $Q(\cdot)$ the Q-Former. For each triplet (r_i, m_i, t_i), the reference image r_i and modification text m_i are jointly processed to produce a composed query embedding, while the target image t_i is encoded into a target representation:

$$q_i = Q(E_{\text{img}}(r_i), m_i), \qquad v_i = Q(E_{\text{img}}(t_i)). \quad (1)$$

Here, q_i denotes the final text-side [CLS] embedding produced from the joint encoding of r_i and m_i, which serves as the composed query representation. In contrast, $v_i = \{v_i(l)\}_{l=1}^{N_q}$ denotes the set of N_q target-side query tokens output by the Q-Former for t_i. Thus, q_i captures both the identity cues of r_i and the appearance modification specified by m_i, while v_i provides fine-grained target-side token features for relevance scoring. Given the query pair (r_i, m_i), the model is trained so that q_i is highly compatible with the corresponding target token set v_i while remaining distinguishable from other candidate targets in the mini-batch.

Following FAFA [7], we use the FDA loss ($\mathcal{L}_{\text{FDA}}$) as the base retrieval objective. In addition, we introduce a Compositional Preference Loss to provide additional supervision on the relational structure among samples within a mini-batch. The detailed formulation of this loss is described in the next subsection.

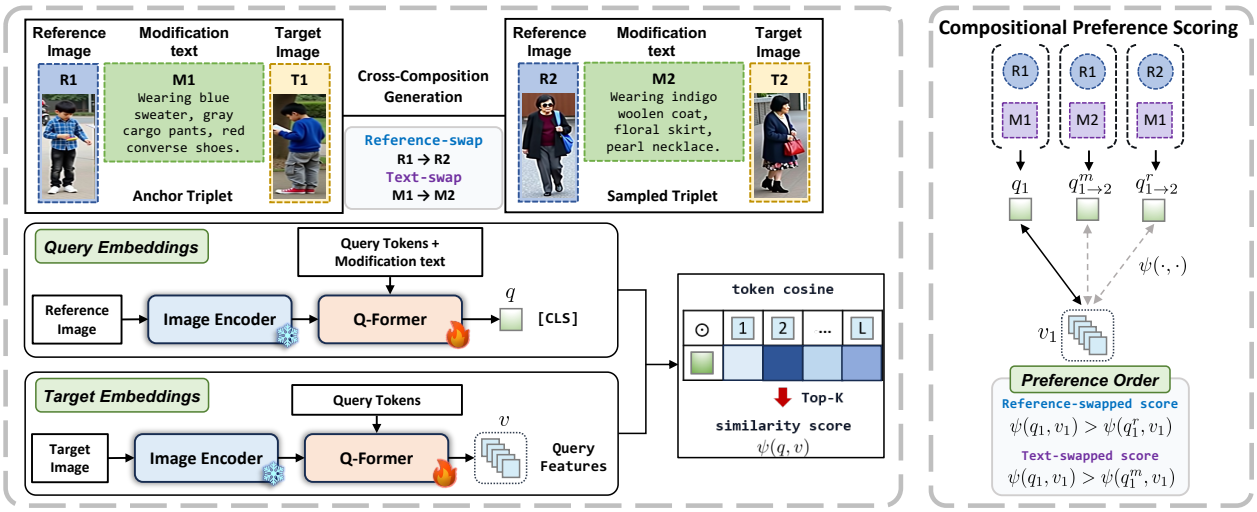

Figure 2. Overview of the COPER framework. COPER encodes the reference image, modification text, and target image with Q-Former, constructs cross-compositions within each mini-batch, and is trained with fine-grained alignment and Compositional Preference supervision.

3.3. Compositional Preference Loss

While the FDA loss encourages fine-grained alignment between a composed query and its target image, it does not explicitly enforce sensitivity to which component of the query has changed. To provide such additional supervision, we introduce a Compositional Preference Loss based on mismatched composition mining.

First, for each sample i in a mini-batch, we randomly select another sample $j \neq i$ from the same batch and construct two cross-composed queries by replacing either the modification text or the reference image:

$$q^m_{i\to j} = Q(E_{\text{img}}(r_i), m_j), \tag{2}$$

$$q^r_{i\to j} = Q(E_{\text{img}}(r_j), m_i). \tag{3}$$

Here, $q^m_{i\to j}$ keeps the reference image fixed while replacing the modification text, whereas $q^r_{i\to j}$ keeps the modification text fixed while replaces the reference image. These cross-composed queries remain semantically close to the original composition, but they should be less compatible with the corresponding target image.

Next, we measure the compatibility between a query embedding q and a target representation v using token-level cosine similarities followed by top-K aggregation:

$$\psi(q,v) = \frac{1}{K}\sum_{k=1}^{K} \text{TopK}_k\left(\left\{\frac{q^\top v_i(l)}{\|q\|\|v_i(l)\|}\right\}_{l=1}^{N_q}\right), \tag{4}$$

where $v(l)$ denotes the l-th target token and $\text{TopK}_k(\cdot)$ denotes the k-th largest value. Based on this measure, the compatibility score of the original composition is defined as

$$s^+_i = \psi(q_i, v_i). \tag{5}$$

We then compute the compatibility scores of the two cross-composed queries with respect to the corresponding target:

$$s^{x,i}_{i\to j} = \psi(q^x_{i\to j}, v_i), \qquad x \in \{m, r\}. \tag{6}$$

Finally, we encourage the original composition to be more compatible with the target image than either cross-composed variants:

$$\mathcal{L}^x_{\text{pref}} = -\log\sigma\left(\frac{s^+_i - s^{x,i}_{i\to j}}{\tau}\right), \qquad x \in \{m, r\}, \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function and τ is a temperature parameter. The final Compositional Preference Loss is defined as

$$\mathcal{L}_{\text{CP}} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathcal{L}^m_{\text{pref}} + \mathcal{L}^r_{\text{pref}}\right). \tag{8}$$

This objective encourages the model to remain sensitive to changes in either the modification text or the reference identity, enabling finer-grained compositional reasoning. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\text{FDA}} + \lambda\mathcal{L}_{\text{CP}}, \tag{9}$$

where $\mathcal{L}_{\text{FDA}}$ denotes the base FDA loss and λ controls the contribution of the proposed compositional preference loss.

4. Experiments

4.1. Datasets

We follow the experimental protocol of FAFA [7] and use separate datasets for training and evaluation. For training,

Table 1. Comparison of our method with representative approaches from related retrieval settings, including ZSCIR, CIR, ZSCPR, and CPR. "Combination" indicates that both the reference image and the modification text are used as the query. "PR" denotes the journal *Pattern Recognition*. The best overall results are highlighted in bold.

Domain	Method	Venue	Pretraining Data	Setting	Rank-1	Rank-5	Rank-10	mAP
ZSCIR	Pic2Word [10]	CVPR'23	CC3M [11]	Combination	21.21	37.15	44.51	29.11
	CoVR-BLIP [13]	AAAI'24	WebVid-CoVR [13]	Combination	26.75	47.68	56.36	36.49
	LinCIR (ViT-G) [2]	CVPR'24	–	Combination	23.93	44.46	53.18	33.95
CIR	CaLa [3]	SIGIR'24	SynCPR [7]	Combination	39.33	60.85	68.66	49.29
	SPRC [15]	ICLR'24	SynCPR [7]	Combination	42.27	61.81	69.35	51.62
ZSCPR	Word4Per [8]	ARXIV	CUHK-PEDES [6]	Combination	44.23	64.76	72.17	54.44
	I2ID [16]	PR'26	CUHK-PEDES [6]	Combination	46.41	66.49	73.52	55.53
CPR	FAFA [7]	NeurIPS'25	SynCPR [7]	Combination	45.46	66.44	73.30	55.15
	COPER(Ours)	-	SynCPR [7]	Combination	**47.64**	**67.85**	**74.98**	**56.95**

we use the synthetic SynCPR dataset [7]. Each sample is constructed via an automatic multimodal data synthesis pipeline and consists of a reference image, a relative caption describing the modification, and the corresponding target image. SynCPR provides large-scale supervision with diverse identities, clothing variations, poses, and scenes, thereby enabling fine-grained alignment between visual and textual cues.

For evaluation, we use the ITCPR benchmark from Word4Per [8], a manually annotated test set introduced for CPR. ITCPR is designed to reflect realistic retrieval scenarios where both a reference image and a relative textual description are available at query time.

4.2. Implementation Details

We train COPER on SynCPR for 10 epochs using AdamW, with an initial learning rate of 4e-6 and a global batch size of 256. All experiments are conducted on a single NVIDIA B200 GPU. The model adopts EVA-CLIP-g [12] as the image encoder with an input resolution of 224×224, along with a Q-Former containing 32 learnable query tokens. For optimization, we set the weight of the proposed Compositional Preference Loss to $\lambda = 1.0$ and the temperature parameter to $\tau = 0.07$.

4.3. Comparison with State-of-the-Art

We compare our method with representative approaches from related retrieval settings, including ZSCIR, CIR, ZSCPR, and CPR, as summarized in Table 1. General composed retrieval methods, including both zero-shot and supervised CIR approaches, remain clearly inferior to CPR-oriented methods, suggesting that person retrieval requires stronger identity-aware compositional reasoning. At the same time, zero-shot CPR methods such as Word4Per and I2ID show competitive performance, highlighting the strong transferability of pretrained vision-language models to this task. Our method achieves the best overall performance, surpassing FAFA by $+2.18\%$ points in Rank-1, $+1.41\%$ points in Rank-5, $+1.68\%$ points in Rank-10, and $+1.80\%$ points in mAP, demonstrating the effectiveness of our approach for CPR.

Table 2. Ablation study on the effect of the proposed Compositional Preference loss $\mathcal{L}_{\text{CP}}$.

Method	Rank-1	Rank-5	Rank-10	mAP
w/o $\mathcal{L}_{\text{CP}}$	46.00	65.99	73.12	55.32
w/ $\mathcal{L}_{\text{CP}}$	47.64 (+1.64)	67.85 (+1.86)	74.98 (+1.86)	56.95 (+1.63)

4.4. Ablation Study

To verify the effectiveness of the proposed Compositional Preference Loss, we compare the full model with a variant trained without $\mathcal{L}_{\text{CP}}$ on ITCPR. As shown in Table 2, removing $\mathcal{L}_{\text{CP}}$ leads to consistent drops across all metrics, confirming its role in learning identity-aware compositional relations.

5. Conclusion

In this paper, we address composed person retrieval by explicitly modeling its identity-sensitive and composition-aware nature. We propose a straightforward framework based on mismatched composition mining and a Compositional Preference Loss, which improves the discrimination between valid and corrupted compositions. We believe this framework can serve as a useful step toward multimodal person retrieval in challenging surveillance scenarios, including aerial and cross-view settings.

References

[1] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15338–15347, 2023. 2

[2] Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. Language-only training of zero-shot composed image retrieval. In *CVPR*, pages 13225–13234, 2024. 2, 4

[3] Xintong Jiang, Yaxiong Wang, Mengjian Li, Yujiao Wu, Bingwen Hu, and Xueming Qian. Cala: Complementary association learning for augmenting comoposed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2177–2187, 2024. 4

[4] Khadija Khaldi, Vuong D. Nguyen, Pranav Mantini, and Shishir Shah. Unsupervised person re-identification in aerial imagery. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 260–269, 2024. 1

[5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[6] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, pages 1970–1979, 2017. 4

[7] Delong Liu, Haiwen Li, Zhaohui Hou, Zhicheng Zhao, Fei Su, and Yuan Dong. Automatic synthetic data and fine-grained adaptive feature alignment for composed person retrieval. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 1, 2, 3, 4

[8] Delong Liu, Haiwen Li, Zhaohui Hou, Zhicheng Zhao, Fei Su, and Yuan Dong. Automatic synthetic data and fine-grained adaptive feature alignment for composed person retrieval. *arXiv preprint arXiv:2311.16515*, 2023. 1, 2, 4

[9] Huy Nguyen, Kien Nguyen, Sridha Sridharan, and Clinton Fookes. Aerial-ground person re-id. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 2585–2590, 2023. 1

[10] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *CVPR*, pages 19305–19314, 2023. 2, 4

[11] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 4

[12] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 4

[13] Lucas Ventura, Antoine Yang, Cordelia Schmid, and G"ul Varol. Covr: Learning composed video retrieval from web video captions. In *AAAI*, pages 5270–5279, 2024. 4

[14] Lei Wang, Quan Zhang, Junyang Qiu, and Jianhuang Lai. Rotation exploration transformer for aerial person re-identification. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2024. 1

[15] Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, Chun-Mei Feng, et al. Sentence-level prompts benefit composed image retrieval. In *ICLR*, 2024. 4

[16] Guo Yu, Di Wang, Chengwei Yan, Feng Yan, Nan Luo, Yifeng Wang, and Quan Wang. I2id: Disentangling identity features via synchronized masking for zero-shot composed person retrieval. *Pattern Recognition*, 179:113654, 2026. 1, 2, 4

[17] Shizhou Zhang, Qi Zhang, Yifei Yang, Xing Wei, Peng Wang, Bingliang Jiao, and Yanning Zhang. Person re-identification in aerial imagery. *IEEE Transactions on Multimedia*, 23:281–291, 2021. 1