Divide and Align: Disentangled Vision-Language Learning for Text-Based Person Anomaly Search

Alex Ergasti[1], Tomaso Fontanini[1], Claudio Ferrari[2], Massimo Bertozzi[1], and Andrea Prati[1]

[1] University of Parma, Department of Architecture and Engineering, Parma, Italy
{alex.ergasti, tomaso.fontanini, massimo.bertozzi, andrea.prati}@unipr.it
[2] Univerisity of Florence, Department of Information Engineering, Firenze, Italy
claudio.ferrari@unifi.it

Abstract. Text-based person anomaly search (TPAS) refers to the task of retrieving people exhibiting normal or anomalous behaviors from natural language descriptions. Existing TPAS models often learn a single joint embedding where appearance, action, and background information are entangled, causing over-reliance on identity cues, poor alignment for action-centric queries, and limited semantic connection between actions and places where they occur. To address these issues, we propose Semantic Decoupled Alignment (SeDA), a disentangled vision–language retrieval framework that explicitly factorizes both visual and textual representations into appearance, action, and background components. SeDA introduces Semantic Token Projection, which derives three semantic queries from the global [CLS] token, softly aggregates modality tokens relevant to each factor, and recomposes the resulting factor tokens into a compact retrieval embedding. To enforce factor-specific semantics, we decompose each caption into appearance/action/background sub-captions and supervise the corresponding tokens with a Feature Decoupling Loss, combined with contrastive and image–text matching objectives. On the Person Anomaly Benchmark (1M pairs), SeDA achieves 86.45% R@1 (+1.52 over SOTA), improves average multi-weather R@1 by +2.43, and gains +3.74 R@1 under out-of-distribution evaluation. https://github.com/ErgastiAlex/SeDA

1 Introduction

The task of retrieving individuals using text descriptions, namely *text-based person search* (TBPS) [8], is a recent yet particularly relevant task in the current landscape of information retrieval applications. Compared to image-based retrieval, TBPS removes the requirement for visual input and leverages natural language, which is more intuitive and practical for deployment in real-world applications. Nevertheless, existing benchmarks, such as [8], [3], and [30], predominantly focus on personal appearance (*e.g.,* clothing, hairstyle, accessory), neglecting other semantically rich factors like actions and environmental context. The latter plays a fundamental role in disambiguating whenever the generic appearance clues become too ambiguous across individuals.

Recently, Yang *et al.* [23] argued that models trained on appearance-centric datasets struggle to be applicable in real world scenarios where people appear similar; thus, searching for specific behaviors or places, rather than focusing solely on static appearance cues, becomes critical. For this reason, they proposed a new task called *text-based person anomaly search* (TPAS) paired with a novel dataset (*Pedestrian Anomaly Behavior*, PAB) that includes highly detailed descriptions. More in detail, TPAS requires retrieving individuals performing both normal and anomalous actions, greatly enhancing the semantic complexity of standard TBPS. Additionally, they also designed a baseline model (*Cross-Modal Pose-aware*, CMP) by integrating pose information and hard negative mining. Nevertheless, CMP neither provided an explicit mechanism to structurally incorporate action information into the learned representation nor adequately modeled the contextual cues in textual descriptions. As a result, the learned representations remain highly entangled, mixing appearance, action, and contextual information within shared feature dimensions, preventing the model from isolating the semantic factors described in textual queries. Consequently, cross-modal alignment becomes less discriminative, leading to reduced retrieval precision and poor generalization in TPAS.

In this paper, we aim to address the entanglement problem by proposing a novel architecture called SeDA for TPAS. Our method explicitly decouples appearance, action, and background components within both visual and textual representations. Specifically, we refer to disentangled semantic components as feature subspaces associated with distinct semantic factors, which are meant to encode complementary, non-overlapping information. In this way, each component focuses on its assigned concept while reducing interference with the others. To achieve this, we design a *Semantic Token Projection* (STP) module. STP first extracts a set of disentangled semantic components and subsequently recombines them into a compact and discriminative representation. Doing so preserves semantic specificity while maintaining a unified embedding space suitable for cross-modal retrieval. To enforce the semantic separation between each component, a *Feature Decoupling Loss* (FDL) is introduced. This objective leverages a large language model (LLM) to extract semantically decomposed captions from the original descriptions, which are then used during training to promote structured cross-modal representations. Overall, our contributions can be summarized as:

- **Explicit semantic decoupling for TPAS.** We propose SeDA, a novel architecture that explicitly decouples appearance, action, and background factors in both visual and textual modalities, addressing the representation entanglement issue in existing TPAS approaches. By structurally separating and recombining semantic factors, our approach reveals that explicitly splitting the semantic information into precise semantic categories improves cross-modal alignment and retrieval robustness in the challenging TPAS setting, particularly for action-centric and anomalous queries.
- **Semantic Token Projection (STP).** We introduce STP, a module that produces semantically disentangled components and recombines them through

similarity-aware projection and aggregation, enhancing semantic consistency and interaction across textual-visual modalities.
– **Feature Decoupling Loss (FDL).** We design FDL, a loss that leverages LLM-based caption decomposition to enforce semantic separation during training and guide structured modality-specific alignment.

2 Related Work

Text-Based Person Search. Text-based person search (TBPS) was first introduced by Li et al. [8] alongside the CUHK-PEDES dataset. By bridging visual perception and linguistic understanding, TBPS opened a new research direction at the intersection of computer vision and natural language processing. Following the release of CUHK-PEDES, several additional datasets have been proposed to further advance research in this area, including ICFG-PEDES [3] and RSTPReID [30]. These datasets aim to address the limitations of earlier benchmarks by providing more diverse identities, richer textual annotations, and more challenging visual conditions, thereby promoting the development of more robust and generalizable TBPS models. In addition to the aforementioned datasets, recent efforts have introduced large-scale synthetic datasets to further enhance backbone pretraining for TBPS. The Multi-Attribute and Language Search (MALS) dataset [24] is a fully synthetic benchmark in which both person images and their corresponding textual descriptions are generated. In contrast, LUPerson-MLLM [15] is constructed from real-world person images, while the associated captions are automatically generated using Qwen [1]. These synthetic datasets provide scalable and diverse training data, enabling improved vision–language representation learning and alleviating the high cost of manual annotation. Early TBPS architectures [16, 18, 19, 26, 28–30] were primarily based on dual-branch frameworks trained from scratch. In these approaches, visual and textual modalities are processed by separate networks, and cross-modal alignment is achieved through carefully designed loss functions or other mechanisms. More recently, Yan *et al*. [20] were the first to systematically investigate the use of pretrained vision–language models, specifically CLIP [12], for the TBPS task. By leveraging CLIP large-scale pretraining on image–text pairs, their work demonstrated significant improvements in cross-modal representation alignment and retrieval performance. This paradigm shift has sparked growing interest in CLIP-based TBPS models, and numerous subsequent works [2, 4, 5, 9] have explored different adaptation strategies, architectural modifications, and fine-tuning schemes built upon CLIP [12], BLIP [6], and ALBEF [7].

Text-Based Person Anomaly Search. In traditional TBPS settings, pedestrians typically exhibit very limited action variability, with most individuals performing similar actions, predominantly walking. This lack of action diversity restricts the applicability of TBPS models in scenarios where behavioral cues are crucial. To address this limitation, Text-based Person Anomaly Search (TPAS) has recently been proposed as a more challenging and practically-relevant task.

In this context, Yang *et al.* [23] introduced the Person Anomaly Benchmark (PAB) dataset, which significantly extends the scope of text-based retrieval to include anomalous human behaviors. The PAB dataset comprises 1,015,583 image–text pairs annotated with fine-grained action descriptions. Unlike conventional TBPS datasets, PAB includes 1,000 distinct *normal* actions, such as walking and running, as well as 1,600 *anomalous* actions, including lying on the ground, being hit, or falling. By explicitly modeling abnormal behaviors, the PAB dataset enables the study of text-based retrieval under more realistic surveillance conditions, where identifying unusual or suspicious activities is of primary importance.

Although the introduction of the PAB dataset [23] has substantially broadened text-based retrieval to include fine-grained normal and anomalous behaviors, TPAS remains less explored than conventional text-based person retrieval. Most existing approaches are still built by directly adapting TBPS pipelines and CLIP-based vision-language models, typically learning a single joint embedding in which appearance, action, and background cues are implicitly mixed. Such entanglement can encourage shortcut learning (*e.g.*, over-reliance on identity-like or contextual cues) and weaken alignment for action-centric queries, which are central to anomaly search. These observations highlight the need for structured representations that separate and align complementary semantic factors across modalities, motivating our disentangled formulation in the proposed framework.

3 Methodology

In SeDA, whose architecture is shown in Fig. 1, we propose a method to decouple appearance, action, and background in order to address the entanglement problem inherent in TPAS. Indeed, current solutions tend to focus disproportionately on person identity while neglecting other distinctive features present in both captions and images, such as performed actions and background context.

To address this limitation, we introduce a novel formulation for computing the [CLS] token called STP (Sec. 3.2), which is designed to disentangle these semantic attributes across visual and textual modalities. This design prevents the model from over-prioritizing identity-related information and encourages a more balanced representation of complementary factors. Furthermore, we propose a novel loss called FDL (Sec. 3.3) that explicitly supervises appearance, action, and background components via semantically decomposed captions.

3.1 Baseline Architecture.

We start from the CMP framework [23] which is based on X-VLM [27]. The model consists of three main components: a pose-aware image encoder $\mathcal{E}_I$, a text encoder $\mathcal{E}_T$, and a cross-modal encoder $\mathcal{E}_C$. The image encoder is based on a Swin Transformer (Swin-B) [10] and takes as input an RGB image I together with its corresponding pose map P. The latter is obtained by rendering a set of detected body joints into a 2D image. The two inputs are jointly processed to

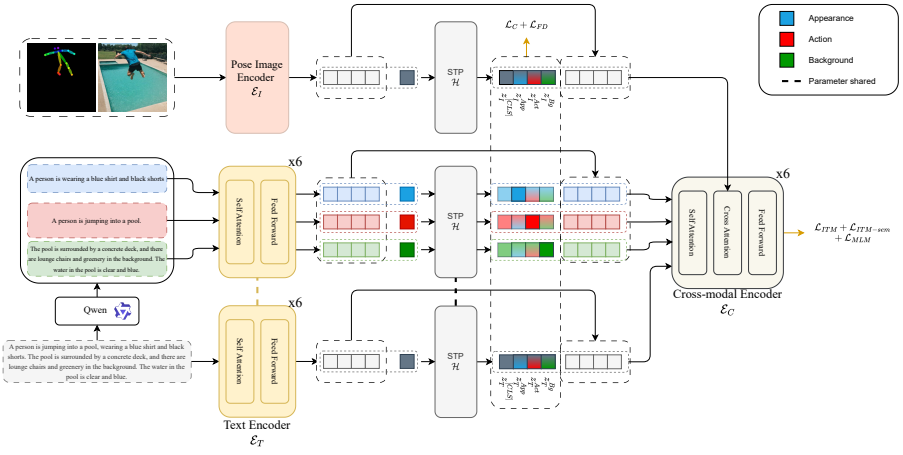

Fig. 1: Overall framework of SeDA. Given an input image and its corresponding caption, the caption is first semantically split into appearance, action, and background components using Qwen [21] (only required during training). Visual and textual features are then extracted by modality-specific encoders. The global [CLS] token is processed by the proposed STP module to obtain disentangled semantic tokens. These tokens are explicitly aligned across modalities through contrastive loss $\mathcal{L}_C$, image-text matching $\mathcal{L}_{ITM}$ and $\mathcal{L}_{ITM-sem}$, masked language modeling loss $\mathcal{L}_{MLM}$, and the proposed Feature Decoupling Loss $\mathcal{L}_{FD}$.

produce a sequence of visual embeddings: $f_I = \mathcal{E}_I(I, P)$. The pose map explicitly injects structural human-body information into the visual representation to better capture pose-dependent, action-related appearance features. For textual processing, we employ a BERT-based architecture. The first six transformer layers are used as the text encoder $\mathcal{E}_T$ to extract textual embeddings from a caption T, $f_T = \mathcal{E}_T(T)$, while the remaining six layers act as the cross-modal encoder $\mathcal{E}_C$, where visual and textual representations interact through cross-attention mechanisms to refine multimodal alignment.

3.2 Semantic Token Projection

To extract structured and decoupled semantic tokens from both image and text modalities, we introduce the STP module $\mathcal{H}(\cdot)$, illustrated in Fig. 2. Given the output of a modality encoder $\mathcal{E}_M$ (where $M \in \{I, T\}$ denotes image or text), we indicate the sequence of token embeddings as $f_M = \{x_M^{[CLS]}, x_M^1, x_M^2, \ldots, x_M^L\}$, where $x_M^{[CLS]} \in \mathbb{R}^d$ is the global [CLS] token, d being the embedding dimension, and $\{x_M^i\}_{i=1}^L$ are the remaining token embeddings. Initially, we generate three semantic queries corresponding to appearance, action, and background directly from the global representation:

$$[q_M^{App}, q_M^{Act}, q_M^{Bg}] = \text{Split}\big(x_M^{[CLS]} W_d^M\big), \tag{1}$$

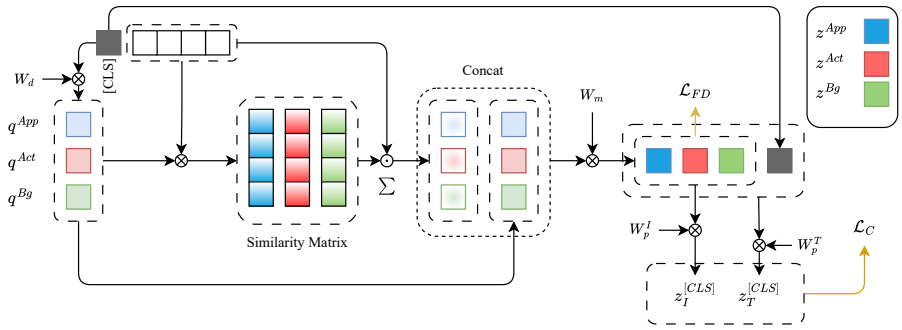

Fig. 2: STP architecture. The global [CLS] token is projected through W_d into three semantic queries, $\mathbf{q} = [q^{App}, q^{Act}, q^{Bg}]$. These queries are used to compute a similarity matrix with the embeddings. The resulting similarity scores are used to obtain weighted sums of the embeddings, which are concatenated with the corresponding queries. The concatenated representations are then projected through W_m obtaining the semantic features z^{App}, z^{Act}, and z^{Bg}. Finally, the updated [CLS] token is obtained via the projection layer W_p. For the image branch, it is computed from the concatenation of the three semantic features. For the text branch, it is derived from the concatenation of the three semantic features and the original [CLS] token.

where $W_d^M \in \mathbb{R}^{d\times 3d}$ is a learnable linear projection for modality M, and Split$(\cdot)$ divides the projected feature into three equal parts. Then, for each semantic factor $k \in \{App, Act, Bg\}$, we compute a similarity score between each token embedding $\{x_M^i\}_{i=1}^L$ and the corresponding semantic query:

$$\alpha_M^{i,k} = \sigma\left(x_M^i \cdot q_M^k\right), \tag{2}$$

where $\sigma(\cdot)$ denotes the sigmoid function. These scores act as soft attention weights over the token sequence to emphasize modality-specific features relevant to each semantic factor. The weighted aggregation is computed as:

$$\tilde{x}_M^k = \sum_{i=1}^{L} (\alpha_M^{i,k}\, x_M^i), \tag{3}$$

Then, each disentangled semantic token is obtained by concatenating the semantic query and its aggregated feature, followed by a modality-specific projection:

$$z_M^k = \left[\, q_M^k \,\|\, \tilde{x}_M^k \,\right] W_m^M, \tag{4}$$

where $W_m^M \in \mathbb{R}^{2d\times d}$ is a learnable projection layer and $\|$ denotes concatenation. In this way, we obtain three disentangled tokens (*i.e.*, appearance, action, and background), which are used to compute a generalized [CLS] representation without over-prioritizing any single component. Finally, the generalized [CLS] token is obtained by concatenating the semantic tokens and projecting them:

$$z_T^{[CLS]} = \left[x_T^{[CLS]} \,\|\, z_T^{App} \,\|\, z_T^{Act} \,\|\, z_T^{Bg} \right] W_p^T, \quad z_I^{[CLS]} = \left[z_I^{App} \,\|\, z_I^{Act} \,\|\, z_I^{Bg} \right] W_p^I, \tag{5}$$

where $W_p^T \in \mathbb{R}^{4d\times d}$ and $W_p^I \in \mathbb{R}^{3d\times d}$. As in [23], we exclude $x_I^{[CLS]}$ when computing $z_I^{[CLS]}$ because, in the Swin-B Transformer backbone used in CMP, the image [CLS] token is defined as the global pooling of the image embeddings. In our framework, this global token is further used to generate the semantic queries that produce z_I^{App}, z_I^{Act}, and z_I^{Bg}. Consequently, its semantic content is implicitly redistributed into these disentangled components. Concatenating $x_I^{[CLS]}$ again would therefore reintroduce mixed, globally entangled information that overlaps with the semantic tokens, undermining the intended decoupling. In contrast, we retain $x_T^{[CLS]}$ since, in the BERT-based text encoder, it encodes complementary sentence-level semantics that are not redundantly reconstructed by the textual semantic components. Thus, for each modality $M \in \{I, T\}$, we obtain the structured token set, composed of a general, fused [CLS] token $z_M^{[CLS]}$ and 3 semantically specific tokens:

$$\{z_M^{[CLS]}, z_M^{App}, z_M^{Act}, z_M^{Bg}\} = \mathcal{H}(f_M). \tag{6}$$

3.3 Feature Decoupling Loss

Without explicit supervision, there is no guarantee that the network will effectively separate the actual semantic factors into the distinct tokens z^{App}, z^{Act} and z^{Bg}. Therefore, to enforce disentanglement, we introduce FDL, which explicitly learns to associate each token with its corresponding semantic attribute.

Caption processing. To structurally disentangle semantic factors in the textual modality, we decompose each caption T into three separate descriptions corresponding to appearance, action, and background, which will be used as anchors. We use Qwen3-8B [21] for this purpose. The model is prompted as follows:

Given this caption, divide it into three separate descriptions: 1) appearance description, 2) action description, and 3) background description. Do not add any information that is not in the original caption. Provide the output in the following JSON format: "appearance": "...", "action": "...", "background": "...".

This procedure outputs three captions: T^{App}, T^{Act}, and T^{Bg}. These structured textual components are used to supervise the disentanglement of representation in both textual and visual embeddings. Note that this step is only required during training. At test time, splitting the caption is not required as the model will learn to disentangle the global [CLS] token.

Loss Formulation. Given a caption T and its decomposed components T^{App}, T^{Act}, and T^{Bg}, we encode them using the shared text encoder $\mathcal{E}_T$:

$$f_T = \mathcal{E}_T(T), \quad f_{T^{App}} = \mathcal{E}_T(T^{App}), \quad f_{T^{Act}} = \mathcal{E}_T(T^{Act}), \quad f_{T^{Bg}} = \mathcal{E}_T(T^{Bg}). \tag{7}$$

Similarly, the image and pose inputs are encoded as:

$$f_I = \mathcal{E}_I(I, P). \tag{8}$$

All embeddings are projected through the STP, $\mathcal{H}(\cdot)$. Specifically, for the global text and image features f_T and f_I, we obtain:

$$\{z_T^{[CLS]}, z_T^{App}, z_T^{Act}, z_T^{Bg}\} = \mathcal{H}(f_T), \quad \{z_I^{[CLS]}, z_I^{App}, z_I^{Act}, z_I^{Bg}\} = \mathcal{H}(f_I) \tag{9}$$

Similarly, we apply $\mathcal{H}(\cdot)$ to the decomposed caption features $f_{T^{App}}$, $f_{T^{Act}}$, and $f_{T^{Bg}}$, and for each branch, we retain only the semantic token corresponding to its factor, yielding $\{z_{T^{App}}^{App}, z_{T^{Act}}^{Act}, z_{T^{Bg}}^{Bg}\}$. For each semantic factor $k \in \{App, Act, Bg\}$, we align the image token z_I^k and the text token z_T^k with the corresponding semantic token $z_{T^k}^k$ from the decomposed caption. FDL is then defined as:

$$\mathcal{L}_{FD} = \sum_{k\in\{App,Act,Bg\}} \left(-\log \frac{\exp\left(\text{sim}(z_I^k, z_{T^k}^k)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(z_{I_j}^k, z_{T^k}^k)/\tau\right)} - \log \frac{\exp\left(\text{sim}(z_T^k, z_{T^k}^k)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(z_{T_j}^k, z_{T^k}^k)/\tau\right)} \right), \tag{10}$$

with $\tau = 0.07$. As a result of this loss, each token is encouraged to encode information specific to its designated factor, *i.e.*, appearance, action, or background, while minimizing interference from other factors. In particular, when $\mathcal{H}(\cdot)$ is applied to a decomposed caption feature (*e.g.*, $f_{T^{App}}$), the discriminative information should be concentrated in the corresponding semantic token (*e.g.*, $z_{T^{App}}^{App}$), whereas the remaining tokens are implicitly discouraged from capturing unrelated content.

3.4 Image-Text Matching loss

Given an image-text pair (I, T) with their embeddings and [CLS] tokens, the cross-modal encoder first processes the text through self-attention and feed-forward layers, then uses it as the query in cross-attention, while the image features serve as keys and values. The updated [CLS] token produced by the cross-modal encoder is then projected to a scalar probability $\hat{p}(I, T)$, representing the predicted likelihood that the image and text are semantically matched. This is optimized against the ground-truth label $p(I, T)$ using a binary cross-entropy:

$$\mathcal{L}_{ITM} = -\mathbb{E}\left[p(I,T)\log\hat{p}(I,T) + (1-p(I,T))\log(1-\hat{p}(I,T))\right]. \tag{11}$$

where $p(I, T) = 1$ for positive pairs and 0 for negative pairs.

In our framework, this loss is applied not only to the original pair (I, T) but also to the semantic-specific pairs (I, T^{App}), (I, T^{Act}), and (I, T^{Bg}). We refer to this extension as the semantic-aware image-text matching loss, denoted by $\mathcal{L}_{ITM\text{-}sem}$, defined as the sum of the ITM losses computed for each component:

$$\mathcal{L}_{ITM\text{-}sem} = \sum_{k\in\{App,Act,Bg\}} \mathcal{L}_{ITM}^k. \tag{12}$$

By enforcing matching at the appearance, action, and background levels, it promotes fine-grained cross-modal alignment across disentangled semantic factors.

Finally, we adopt in-batch hard negative mining by sampling mismatched image-text pairs based on their similarity scores, encouraging the model to focus on more challenging examples. For the global pair (I, T), sampling probabilities are computed from the similarity between normalized $z^{[CLS]}$ representations, making harder negatives more likely to be selected. We further extend this strategy to the semantic-specific pairs (I, T^{App}), (I, T^{Act}), and (I, T^{Bg}), enabling factor-level hard negative selection and enhancing discriminative learning.

3.5 Cross-modal Objectives

To align structured visual and textual representations, we adopt a set of cross-modal objectives that promote both global and semantic-specific consistency between modalities. Specifically, we employ a contrastive loss $\mathcal{L}_C$ between $z_T^{[CLS]}$ and $z_I^{[CLS]}$, a masked language modeling loss $\mathcal{L}_{MLM}$ [5], and the identity-based hard negative mining [23]. The final loss is then composed as:

$$\mathcal{L} = \mathcal{L}_{FD} + \mathcal{L}_C + \mathcal{L}_{ITM} + \mathcal{L}_{ITM-sem} + \mathcal{L}_{MLM} \tag{13}$$

4 Experiments

Training Details. We train SeDA on 4 NVIDIA L40S GPUs, using a mini batch size of 22 on each GPU for 30 epochs with a learning rate of $1e-4$, which linearly decreases until $1e-5$, and AdamW [11] as optimizer with a weight decay of 0.01. We started from the weights of X-VLM [27]. All loss terms have equal weights.

Evaluation metrics. We evaluate our model using standard retrieval metrics. Specifically, we report Rank@K, which is reported as R@1, R@5, R@10, and measures the fraction of queries for which at least one relevant result appears among the top-K retrieved items. In addition, we report the mean Average Precision (mAP) to capture the overall ranking quality of the retrieved items.

Datasets. We trained our model on the PAB dataset [23] and evaluated it on its test set. We also tested the model trained on PAB in a multi-weather setting [17,23] and on an out-of-distribution (OOD) set on the UCC dataset [23]. In the multi-weather setting the PAB test set is augmented to include 10 different environmental conditions (*e.g.* wind, rain, snow, etc.) that simulate a realistic smart city scenario. For the OOD evaluation, 5,320 image–text pairs are extracted from UCF-Crime [13], showcasing both normal and anomalous events.

4.1 Quantitative Results

In Table 1, we compare our model with state-of-the-art methods in TBPS and TPAS. Notably, our model outperforms all TBPS and TPAS baselines even when trained with only a small fraction of the available data (0.1M samples).

Method	Data	R@1	R@5	R@10	mAP
MRA [22]	0.1M	70.53	94.69	97.47	81.59
APTM [24]	0.1M	72.14	95.30	97.17	82.78
CAMeL [25]	0.1M	74.30	96.79	98.84	84.20
WoRA [14]	0.1M	74.47	96.82	98.48	84.60
IRRA [5]	0.1M	76.39	97.62	99.14	86.33
CLIP [12]	0.1M	77.60	98.84	99.75	87.35
RaSa [2]	0.1M	80.79	98.89	99.65	89.20
X-VLM [27]	0.1M	81.95	98.84	99.19	89.86
CMP [23]	0.1M	83.06	98.89	99.49	90.41
CMP [23]	1M	84.93	99.09	99.75	91.66
Ours	0.1M	<u>85.14</u>	<u>99.24</u>	<u>99.80</u>	<u>91.83</u>
Ours	1M	**86.45**	**99.44**	**99.85**	**92.59**

Table 1: Comparison with state-of-the-arts models on PAB dataset. Previous results are taken from [23]. Best results in bold, second best underlined.

Furthermore, when trained on the full dataset, SeDA achieves an R@1 of 86.45%, corresponding to a +1.52 improvement over the previous state of the art.

We further analyze the efficiency of SeDA compared with CMP. Although SeDA introduces only a small increase in parameters (+4.0%), it incurs higher training cost (+16% time and +19.4% memory). However, it requires only 1/10 of the training data to surpass CMP. Moreover, since split captions are used only during training, SeDA and CMP have the same inference speed and nearly identical memory usage.

In Table 2, we report the performance of our model under multi-weather conditions. Compared with CMP, our model achieves the best results across all weather settings, improving R@1 by +2.43, mAP by +1.90, and demonstrating stronger robustness and reliability in this challenging setting.

Method	Metric	Normal	Wind	Rain	Snow	R+S	Dark	D+W	D+R	D+S	OverExp	Mean
CMP	**R@1**	84.93	81.24	60.06	63.40	54.85	80.89	77.20	39.03	53.49	76.09	67.12
Ours		**86.45**	**83.42**	**61.78**	**66.18**	**55.46**	**83.82**	**80.49**	**41.15**	**57.69**	**79.02**	**69.55**
CMP	**mAP**	91.66	89.34	72.53	75.74	67.31	89.00	86.55	50.58	66.12	85.56	77.44
Ours		**92.59**	**90.72**	**73.88**	**78.17**	**68.02**	**90.97**	**88.69**	**52.42**	**70.45**	**87.46**	**79.34**

Table 2: Multi-weather comparison between CMP and Ours. Best results in bold. R+S: Rain+Snow, D+W: Dark+Wind, D+R: Dark+Rain, D+S: Dark+Snow.

In Table 3, we report the comparison between CMP and our model on the person retrieval task calculated on PAB, instead of the anomaly person retrieval setting. Our model outperforms the baseline in this task as well, showing that emphasizing fine-grained semantic details does not hinder person retrieval, where appearance cues are primarily important, but in fact enhances retrieval per-

formance. Importantly, person retrieval on PAB appears near-saturated (both methods already achieve very high performance), leaving limited headroom for gains driven purely by appearance. Therefore, the larger improvements observed in the TPAS setting are best interpreted as stemming from better alignment of action and contextual/background cues, which are the key discriminative factors for anomalous behaviors.

Method	R@1	R@5	R@10	mAP
CMP	94.34	99.39	99.85	88.01
Ours	**94.69**	**99.70**	**99.90**	**88.08**

Table 3: Comparison between CMP and SeDA in Identity Search Retrieval on PAB.

In Table 4, we evaluate our model in an out-of-distribution (OOD) setting. Compared with prior TBPS and TPAS methods, SeDA achieves the best performance in all the metrics. Notably, our model surpasses the previous state of the art (CMP) by a substantial margin of +3.74 in R@1, highlighting a significant improvement in top-ranked retrieval under distribution shifts. This experiment validates the generalization capabilities of the proposed model.

Method	R@1	R@5	R@10	mAP
APTM	27.86	40.41	46.77	22.61
IRRA	40.28	57.24	65.98	33.53
CLIP	51.60	68.31	76.43	43.05
X-VLM	52.33	66.73	72.54	40.87
RaSa	54.12	70.32	75.96	39.71
CMP	<u>55.23</u>	<u>71.67</u>	<u>77.99</u>	<u>44.35</u>
Ours	**58.97**	**73.37**	**79.08**	**44.57**

Table 4: Comparison with state-of-the-art methods in OOD setting on UCC dataset. All models trained on 1M images. Best results in bold, second best underlined.

4.2 Qualitative Results

To further analyze the effectiveness of our approach, we visualize the similarity matrix produced within the STP module, computed between the projected semantic queries (q^{App}, q^{Act}, q^{Bg}) derived from the [CLS] token and the corresponding token embeddings of Eq. 2. As shown in Fig. 3, each semantic query exhibits focused similarity responses over modality-specific semantic regions. For example, q_T^{App} shows higher similarity scores for words related to appearance,

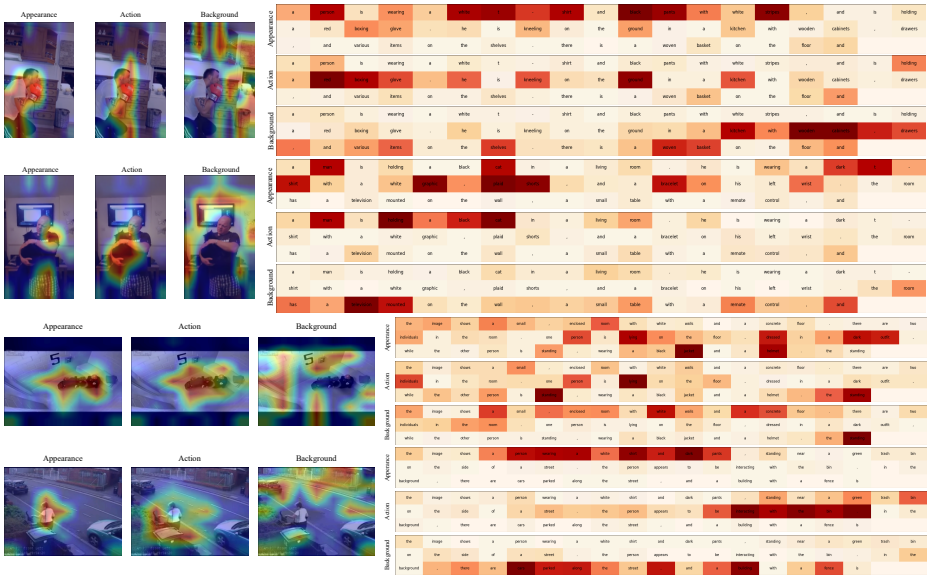

Fig. 3: Visualization of the similarity matrix calculated in both the Image Encoder and the Text Encoder. On the left, there are four examples (first two on PAB and last two on UCC) of the similarity matrix between $q_I^{App}, q_I^{Act}, q_I^{Bg}$ and the image embeddings. On the right, there is the similarity matrix between $q_T^{App}, q_T^{Act}, q_T^{Bg}$ and the text embeddings (captions are capped at 56 words consistently with SOTA methods). As it can be seen, the similarity matrix focuses on the corresponding embeddings, demonstrating the utility of both STP and $\mathcal{L}_{FD}$. Better zoomed in.

while q_I^{App} concentrates on image patches corresponding to appearance cues. Similar behavior is observed for the action and background queries, confirming that the semantic token projection combined with $\mathcal{L}_{FD}$ effectively disentangles modality-specific attributes.

Finally, in Fig. 4, we present four examples of anomaly retrieval comparing our model with CMP. In both cases, CMP fails to retrieve the correct image in the first position, while our model successfully identifies the correct image. In detail, on the left, our model is able to better discriminate the action; on the right, it identifies the correct identity, showing the capability to focus on each part of the caption.

4.3 Ablation

To validate the design of SeDA, we performed several ablation studies. In Table 5, we analyze the contribution of different variants of the STP module, as well as the roles of $\mathcal{L}_{FD}$ and $\mathcal{L}_{ITM\text{-}sem}$. These ablation experiments are conducted on a reduced subset (0.1M samples) of the PAB dataset.

We first compare three variants of the STP module, all trained with the proposed $\mathcal{L}_{FD}$: (I) replacing the sigmoid with softmax, (II) introducing a cross-

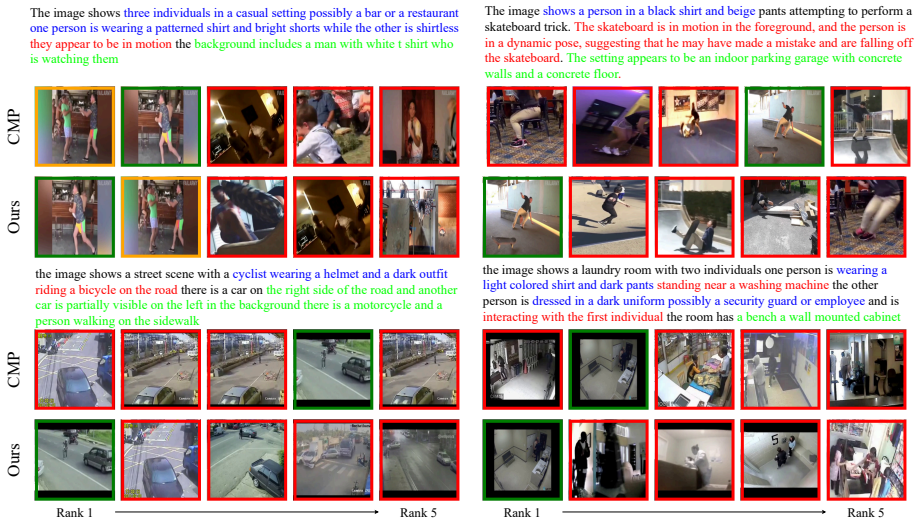

Fig. 4: Comparison of ranking results between CMP and our model on PAB (first row) and UCC (second row). Correct image is highlighted in green, correct identity but wrong action is highlighted in orange, wrong identity in red.

attention module that relies on softmax-based attention, and (III) the proposed STP design. Variants (I) and (II) both underperform compared to (III). We attribute this to the softmax normalization, which, although producing sum-to-one weights, introduces exponential inter-token competition and enforces a winner-take-most behavior. In contrast, the proposed STP module enables cooperative aggregation, allowing multiple relevant regions or words to contribute jointly, thereby improving performance.

Then, in (IV), we remove $\mathcal{L}_{FD}$. The resulting performance drop demonstrates that this loss is crucial for enforcing modality disentanglement. Indeed, without it, the STP module fails to learn properly decoupled representations, and results fall back to being similar to those of the CMP baseline.

Finally, in (V), we add the semantic-specific ITM loss $\mathcal{L}_{ITM\text{-}sem}$, which further improves the results w.r.t. (III).

Finally, Table 6 reports the performance of different combinations of semantic components. First, each individual component already achieves strong performance, indicating that appearance, action, and background cues are all individually informative. However, combining components consistently improves the results, confirming that these cues provide complementary information. In particular, pairwise combinations outperform single components across all metrics. The full combination (STP module) achieves the best overall performance. This demonstrates that our proposed aggregation strategy effectively leverages complementary semantic cues, leading to more discriminative and robust representations.

Exp.	Method	R@1	R@5	R@10	mAP
(0)	CMP (baseline)	83.06	98.89	99.49	90.41
(I)	$\text{STP}_{softmax}$	83.82	99.04	99.65	90.95
(II)	STP_{CA}	83.92	98.94	99.44	90.92
(III)	STP	84.78	99.04	99.59	91.54
(IV)	STP w/o $\mathcal{L}_{FD}$	83.15	99.09	99.39	90.52
(V)	STP + $\mathcal{L}_{ITM-sem}$	**85.14**	**99.24**	**99.80**	**91.83**

Table 5: Ablation study with different versions of the proposed STP module and loss combinations.

Method	Appearance	Action	Background	R@1	R@5	R@10	mAP
(I)	✓	✗	✗	83.67	98.43	98.69	90.64
(II)	✗	✓	✗	83.92	97.42	97.92	90.39
(III)	✗	✗	✓	82.10	97.07	97.52	89.33
(IV)	✓	✓	✗	84.50	98.94	99.34	91.41
(V)	✓	✗	✓	84.38	99.19	99.77	91.40
(VI)	✗	✓	✓	84.88	98.94	99.49	91.55
STP	✓	✓	✓	**85.14**	**99.24**	**99.80**	**91.83**

Table 6: Performance comparison in terms of R@1, R@5, R@10, and mAP of different combinations of semantic components calculated on PAB dataset.

5 Conclusion

In this paper, we introduced SeDA, a disentangled vision–language framework for text-based person anomaly search (TPAS) that tackles the key limitations of existing methods: semantic entanglement between appearance, action, and background, which weakens action-centric alignment and hurts generalization. SeDA combines a Semantic Token Projection (STP) module, designed to produce factor-specific components and then recombine them into a compact retrieval embedding, with a Feature Decoupling Loss (FDL) that leverages LLM-based caption decomposition to supervise each factor with the appropriate textual signal. Extensive experiments on the Person Anomaly Benchmark (PAB) show that SeDA achieves 86.45% R@1 (+1.52 R@1 gain over CMP) when trained on the full data, outperforming prior approaches, while also improving robustness under challenging multi-weather conditions (average +2.43 R@1 over CMP) and under out-of-distribution evaluation on UCC (+3.74 R@1 gain over CMP).

References

1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., Hui, B., Ji, L., Li, M., Lin, J., Lin, R., Liu, D., Liu, G., Lu, C., Lu, K., Ma, J., Men, R., Ren, X., Ren, X., Tan, C., Tan, S., Tu, J., Wang, P., Wang,

S., Wang, W., Wu, S., Xu, B., Xu, J., Yang, A., Yang, H., Yang, J., Yang, S., Yao, Y., Yu, B., Yuan, H., Yuan, Z., Zhang, J., Zhang, X., Zhang, Y., Zhang, Z., Zhou, C., Zhou, J., Zhou, X., Zhu, T.: Qwen technical report (2023), https://arxiv.org/abs/2309.16609

2. Bai, Y., Cao, M., Gao, D., Cao, Z., Chen, C., Fan, Z., Nie, L., Zhang, M.: RaSa: Relation and Sensitivity Aware Representation Learning for Text-based Person Search. In: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. p. 555–563. IJCAI-2023, International Joint Conferences on Artificial Intelligence Organization (Aug 2023). https://doi.org/10.24963/ijcai.2023/62, http://dx.doi.org/10.24963/ijcai.2023/62
3. Ding, Z., Ding, C., Shao, Z., Tao, D.: Semantically self-aligned network for text-to-image part-aware person re-identification. arXiv preprint arXiv:2107.12666 (2021)
4. Ergasti, A., Fontanini, T., Ferrari, C., Bertozzi, M., Prati, A.: Mars: Paying more attention to visual attributes for text-based person search. ACM Trans. Multimedia Comput. Commun. Appl. **21**(10) (Oct 2025). https://doi.org/10.1145/3721482, https://doi.org/10.1145/3721482
5. Jiang, D., Ye, M.: Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2787–2797 (2023)
6. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022)
7. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
8. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X.: Person search with natural language description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
9. Lin, D., Peng, Y.X., Meng, J., Zheng, W.S.: Cross-modal adaptive dual association for text-to-image person retrieval. IEEE Transactions on Multimedia **26**, 6609–6620 (2024). https://doi.org/10.1109/TMM.2024.3355644
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
13. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6479–6488 (2018)
14. Sun, J., Fei, H., Ding, G., Zheng, Z.: From data deluge to data curation: A filtering-wora paradigm for efficient text-based person search. In: Proceedings of the ACM on Web Conference 2025. pp. 2341–2351 (2025)

15. Tan, W., Ding, C., Jiang, J., Wang, F., Zhan, Y., Tao, D.: Harnessing the power of MLLMs for transferable text-to-image person ReID. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17127–17137 (2024)
16. Wang, C., Luo, Z., Lin, Y., Li, S.: Text-based person search via multi-granularity embedding learning. In: IJCAI. pp. 1068–1074 (2021)
17. Wang, T., Zheng, Z., Sun, Y., Yan, C., Yang, Y., Chua, T.S.: Multiple-environment self-adaptive network for aerial-view geo-localization. Pattern Recognition **152**, 110363 (2024)
18. Wang, Z., Fang, Z., Wang, J., Yang, Y.: ViTAA: Visual-textual attributes alignment in person search by natural language. In: European conference on computer vision. pp. 402–420. Springer (2020)
19. Wu, Y., Yan, Z., Han, X., Li, G., Zou, C., Cui, S.: Lapscore: language-guided person search via color reasoning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1624–1633 (2021)
20. Yan, S., Dong, N., Zhang, L., Tang, J.: Clip-driven fine-grained text-image person re-identification. IEEE Transactions on Image Processing **32**, 6032–6046 (2023)
21. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
22. Yang, S., Wang, Y., Li, Y., Zhu, L., Zheng, Z.: Minimizing the pretraining gap: Domain-aligned text-based person retrieval. arXiv preprint arXiv:2507.10195 (2025)
23. Yang, S., Wang, Y., Zhu, L., Zheng, Z.: Beyond walking: A large-scale image-text benchmark for text-based person anomaly search. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11720–11730 (2025)
24. Yang, S., Zhou, Y., Zheng, Z., Wang, Y., Zhu, L., Wu, Y.: Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In: Proceedings of the 31st ACM international conference on multimedia. pp. 4492–4501 (2023)
25. Yu, H., Wen, J., Zheng, Z.: CAMeL: Cross-modality Adaptive Meta-Learning for Text-based Person Retrieval. IEEE Transactions on Information Forensics and Security (2025)
26. Zeng, P., Jing, S., Song, J., Fan, K., Li, X., We, L., Guo, Y.: Relation-aware aggregation network with auxiliary guidance for text-based person search. World Wide Web **25**(4), 1565–1582 (2022)
27. Zeng, Y., Zhang, X., Li, H.: Multi-grained vision language pre-training: Aligning texts with visual concepts. arXiv preprint arXiv:2111.08276 (2021)
28. Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching. In: Proceedings of the European conference on computer vision (ECCV). pp. 686–701 (2018)
29. Zheng, K., Liu, W., Liu, J., Zha, Z.J., Mei, T.: Hierarchical gumbel attention network for text-based person search. In: Proceedings of the 28th ACM international conference on multimedia. pp. 3441–3449 (2020)

30. Zhu, A., Wang, Z., Li, Y., Wan, X., Jin, J., Wang, T., Hu, F., Hua, G.: Dssl: Deep surroundings-person separation learning for text-based person retrieval. In: Proceedings of the 29th ACM international conference on multimedia. pp. 209–217 (2021)