Beyond Walking: A Large-Scale Image-Text Benchmark
for Text-based Person Anomaly Search

Shuyu Yang[1]   Yaxiong Wang[2*]   Li Zhu[1*]   Zhedong Zheng[3*]

[1]Xi'an Jiaotong University   [2]Hefei University of Technology   [3]University of Macau

{ysy653, wangyx15}@stu.xjtu.edu.cn, zhuli@mail.xjtu.edu.cn, zhedongzheng@um.edu.mo
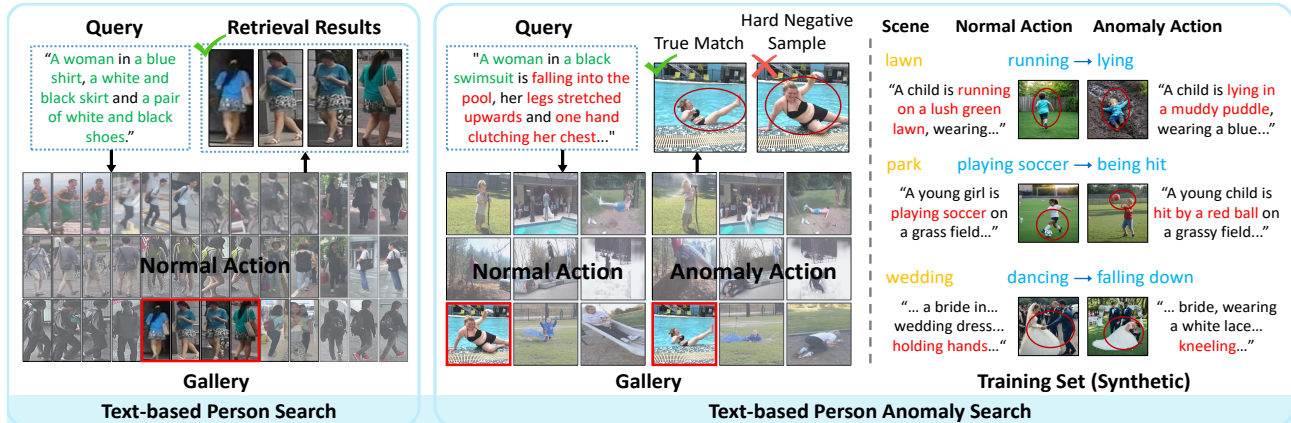
Figure 1. **Comparison of our proposed task, *i.e.*, Text-based Person Anomaly Search (*right*) vs. Traditional Text-Based Person Search (*left*).** Traditional text-based person search primarily focuses on the appearance of individuals and often overlooks action information, if any. In contrast, given the appearance and action description, text-based person anomaly search aims to locate the pedestrian of interest engaged in either normal or abnormal actions from a large pool of candidates. Text-based person anomaly search emphasizes the identification of pedestrian abnormal behaviors, aligning closely with real-world emergency and safety requirements.

Abstract

Text-based person search aims to retrieve specific individuals across camera networks using natural language descriptions. However, current benchmarks often exhibit biases towards common actions like walking or standing, neglecting the critical need for identifying abnormal behaviors in real-world scenarios. To meet such demands, we propose a new task, text-based person anomaly search, locating pedestrians engaged in both routine or anomalous activities via text. To enable the training and evaluation of this new task, we construct a large-scale image-text Pedestrian Anomaly Behavior (PAB) benchmark, featuring a broad spectrum of actions, e.g., running, performing, playing soccer, and the corresponding anomalies, e.g., lying, being hit, and falling of the same identity. The training set of PAB comprises $1,013,605$ synthesized image-text pairs of both normalities and anomalies, while the test set includes $1,978$ real-world image-text pairs. To validate the potential of PAB, we introduce a cross-modal pose-aware framework, which integrates human pose patterns with identity-based hard negative pair sampling. Extensive experiments on the proposed benchmark show that synthetic training data facilitates the fine-grained behavior retrieval, and the proposed pose-aware method arrives at 84.93% recall@1 accuracy, surpassing other competitive methods.

1. Introduction

Text-based person search [25, 68] focuses on retrieving specific individuals from large-scale image databases using natural language descriptions. This capability is particularly valuable for user-interactive applications in smart cities, security systems, and personalized services, where image queries are often unavailable or impractical to obtain. However, current benchmarks for text-based person search suffer from significant limitations in behavioral diversity, primarily due to their bias towards common pedestrian actions. Existing datasets, including CUHK-PEDES [25], ICFG-PEDES [11], and RSTReid [72], predominantly feature routine activities such as walking and standing, failing to represent the full spectrum of real-world behaviors ade-

---

Datasets	Modality	Annotation	Content	#Frames	#Texts	#Action Types	Anomaly:Normal	Data source
CUHK-PEDES [25]	Image, Text	Frame-level Text	Appearance	40,206	80,440	-	0	Collection
ICFG-PEDES [11]	Image, Text	Frame-level Text	Appearance	54,522	54,522	-	0	Collection
RSTPReid [72]	Image, Text	Frame-level Text	Appearance	20,505	41,010	-	0	Collection
UBnormal [1]	Video	Frame-level Tag	Binary Label	236,902	-	22 Anomaly	2:3	Synthesis
ShanghaiTech [33]	Video	Frame-level Tag	Binary Label	317,398	-	11 Anomaly	1:18	Collection
UCF-Crime [45]	Video	Video-level Tag	Binary Label	13,741,393	-	13 Anomaly	$\ll$1:1	Collection
UCA [63]	Video, Text	Video-level Text	Action	13,741,393	23,542	13 Anomaly	$\ll$1:1	Collection
PAB (Ours)	Image, Text	Frame-level Text	**Appearance, Action, Scene**	**1,015,583**	**1,015,583**	**1,600 Anomaly 1,000 Normal**	**3:2**	Synthesis & Collection

Table 1. **Dataset Characteristic Comparison.** We present a comprehensive comparison between our proposed Pedestrian Anomaly Behavior (PAB) benchmark and existing text-based pedestrian search and video anomaly detection datasets in terms of data quality and quantity. Our dataset addresses the long-tail distribution challenge by incorporating a higher number of anomaly cases, while offering frame-level annotations (appearance, action, and scene) and fine-grained action types.

quately (see Fig. 1). This inherent bias significantly restricts the generalizability and practical applicability of models trained on these datasets, particularly in critical scenarios requiring anomaly behavior detection.

Moreover, current anomaly detection methodologies focus on identifying and classifying predefined events, while existing video datasets face three major limitations: (1) limited data volume, typically containing less than 500k frames [27, 32, 35], (2) coarse annotation granularity, often restricted to binary labels (anomaly *vs.* normal) [33, 45, 49], and (3) behavior long-tail distribution, where anomaly key frames are significantly underrepresented compared to normal behaviors [2, 40, 63] (see Table 1). These limitations hinder the development of robust models capable of handling diverse real-world scenarios. While UBnormal [1] addresses multi-class annotations using simulated 3D environments, the gap between virtual scenes and real-world footage limits generalization in practical deployments. Moreover, real-world applications often demand the ability to search for specific behaviors rather than broad, predefined categories, further highlighting the need for more nuanced datasets.

To address the limitations in both fields, we introduce a new task, *i.e.*, text-based person anomaly search. **(1)** This task extends the traditional scope of text-based person search by requiring the identification of pedestrians involved in both routine and anomalous activities. While existing methods could successfully locate a person walking, they often fail to identify the same person lying on the ground or being hit. **(2)** Comparing with the conventional anomaly detection, the proposed task further explores a fine-grained anomaly understanding beyond the binary tags, which is critical for security events tracing and locating, emergency response, and many other security applications. To support this task, we propose the Pedestrian Anomaly Behavior (PAB) benchmark. As summarized in Table 1, the PAB benchmark consists of 1,015,583 image-text pairs, each annotated with detailed textual descriptions of the target pedestrian's appearance, actions, and surrounding scene. The dataset comprehensively spans 1,000 distinct normal action, such as running, performing, and play-

ing soccer, along with 1,600 anomalies like lying, being hit, and falling. This broad coverage ensures diversity in both routine and rare pedestrian activities, while the intentionally elevated anomaly ratio (3:2) enhances its utility for training robust anomaly detection models. To validate the potential of the PAB benchmark, we introduce a Cross-Modal Pose-aware (CMP) framework that integrates human pose patterns with identity-based hard negative pair sampling. This framework leverages the rich pose information to distinguish between normal and anomalous behaviors. Extensive experiments show that synthetic training data facilitate the person anomaly search on the real-world test set. The proposed framework also achieves a substantial improvement. In a summary, our primary contributions are:

- We pioneer the task of text-based person anomaly search that **unifies detection of both routine and anomalous activities** through natural language queries. To address the critical absence of fine-grained anomaly data, we construct the Pedestrian Anomaly Behavior (PAB) benchmark, featuring $1.01M$ image-text pairs spanning $1,600$ anomaly and $1,000$ normal action types.
- Departing from conventional identity-centric person search methods, we propose a Cross-Modal Pose-aware (CMP) framework that integrates text, image, and human pose patterns for representation learning. We also introduce identity-based hard negative mining by perturbing the action to distinguishing subtle behavioral differences.
- Extensive experiments show that (1) synthetic training data of the proposed PAB facilitates fine-grained behavior retrieval in the real-world test set; (2) our full CMP model achieves 84.93% recall@1 accuracy on PAB and 55.23% recall@1 accuracy on the out-of-distribution (OOD) UCC testing, outperforming multiple competitive approaches.

2. Related Work

Text-based Person Retrieval. Text-based pedestrian retrieval incorporates textual queries into large-scale pedestrian retrieval tasks, breaking the limitations of image-based [34, 67, 69] or attribute-based queries [18, 28, 29]. Text-based pedestrian retrieval is generally more challenging than image-based pedestrian retrieval [3, 66, 67, 70]

and general cross-modal retrieval tasks [12, 17, 50, 51, 58, 61] due to its cross-modal and fine-grained nature. The key to this task is to learn the alignment of the image and text [9, 16, 23, 26, 36, 38, 52]. Roughly, prevailing alignment strategies can be divided into cross-modal attention-based [25, 37, 43, 44, 54] and cross-modal non-attention [9, 11, 53, 71] methods. With the development of vision-language pre-training models [24, 39], recent works [5, 19, 22, 44, 57] have improved the robustness of learned features by transferring knowledge from large-scale generic image-text pairs. Inspired by the large language and multimodal models, recent studies attempt to utilize the large model to facilitate this problem [47, 59]. Yang *et al*. [59] presents a large-scale synthetic image-text dataset MALS for pre-training, while Tan *et al*. [47] generate a large-scale dataset to study transferable text-to-image ReID. In this paper, we introduce text-based person anomaly search, focusing on fine-grained behavior retrieval.

Person Anomaly Detection. Existing work on pedestrian anomaly detection typically refers to Video Anomaly Detection (VAD) [6]. The goal of VAD is to detect events in videos that deviate from normal patterns, and it has been studied under various settings: as a one-class classification problem [14, 15, 21, 41, 64] where training only involves normal data; as an unsupervised learning task [64] where anomalies exist in training set but it is unknown which training videos contain them; and as a supervised or weakly supervised problem [1, 45, 64] where training labels indicate anomalous video frames or videos containing anomalies. Most of these works focus on the one-class classification problem. Consequently, many video datasets are limited in terms of the number of videos or the realism of the scenarios. Examples include the UCSD Ped1 and Ped2 [27], Avenue [32], Subway [2], ShanghaiTech Campus [33], and UBnormal [1]. Sultani *et al*. [45] construct a realistic video dataset called UCF-Crime, which annotates 13 pre-defined categories of video anomalies but still falls short in handling complex real-world scenarios. Despite efforts [56, 63] to annotate UCF-Crime videos with text, datasets such as UCA [63] remain constrained (see Table 1). Therefore, we propose a large-scale multi-modal Pedestrian Anomaly Behavior (PAB) dataset that contains large-scale diverse normal and anomalous image-text pairs, capable of addressing more complex multi-modal anomaly behavior retrieval.

3. Person Anomaly Benchmark

3.1. Real-world Test Data Collection

To establish a practical evaluation benchmark, we construct our test set using real-world videos from OOPS! [13].
Anomaly and Anomaly Image Extraction. OOPS! videos come with timestamps indicating when an anomaly or unintentional action begins. This means that the content before

the timestamp depicts normal behavior, while the content after the timestamp shows anomaly behavior. We extract middle frames from the video segments before and after the timestamps as pedestrian normal and anomaly images, respectively, and still face the following issues:

- Noise images, *i.e.*, images that do not contain people or where people occupy a small proportion of the image area. To address this issue, we deploy OpenPose [7] to detect human key points and eliminate undesired images.
- Duplicated images or images with subtle discrepancy. To overcome such noise, we apply ResNet-50 [20] to extract features from the images and calculate the cosine similarity between normal and anomaly images, filtering out pairs with a similarity greater than 0.95.
- False-positive anomaly images, *i.e.*, behaviors in anomaly images are normal. For this problem, we conduct manual verification by three professionals with advanced education in computer science, retaining only those image pairs of certain normal and anomaly behaviors. Through the aforementioned steps, we obtain 989 high-quality image pairs ($1,978$ pedestrian images), which are designated as 1:1 normal and anomaly image pairs for the test set.

Caption Generation and Quality Control. Each OOPS! video is annotated with two types of captions, one (C_n) for normal moment and another (C_a) for anomaly occurring. Directly adopting these captions as image captions does not work well, because most of the captions are short and without detailed descriptions of appearance. Drawing inspiration from the significant progress in the Multi-modal Large Language Model (MLLM) [4, 8], we automatically generate captions for each image, eliminating the need for costly manual annotation. Particularly, we choose Qwen2-VL [4] as Image Captioner. The specific **Instruction** is as follows: "Provide a simple description of the image content within 50 words, including the appearance, attire, and actions of the main figures in the picture. Do not imagine any contents that are not in the image. Do not describe the atmosphere of the image." The captions generated by MLLM can generally provide accurate descriptions of the appearance and actions of people in the images, including detailed information. However, due to inherent limitations of the model, the captions may contain minor errors. Therefore, we conduct **Human Quality Control** on the test set. Specifically, we enlist three professionals with advanced degrees in computer science to manually correct the text data. This measure ensures that the texts accurately match the image content especially for specific normal and anomaly behaviors.

3.2. Large-scale Training Data Generation

To bypass the scarcity of anomaly data, we resort to utilizing the generative model, *e.g*., diffusion model, to synthesize our training data with the following steps:
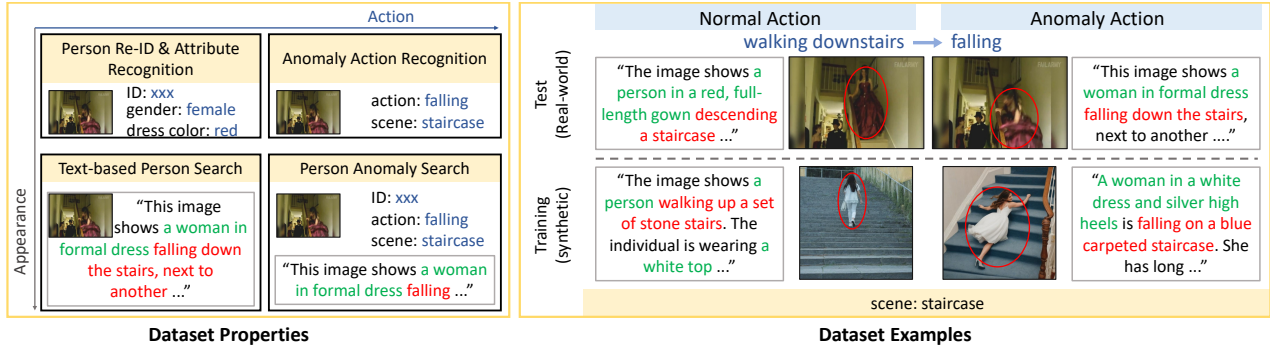**Pedestrian-focused Image Generation.** To ensure the gen-

Figure 2. **Dataset Properties** (*left*). Compared with existing datasets for person re-ID, attribute recognition, text-based person search, and anomaly action recognition, ours contain more detailed action and appearance descriptions for text-based anomaly search. **Dataset Examples** (*right*). Our training set is synthesized, while the test set is collected from real-world videos. We provide similar training samples in terms of normal and anomaly action, facilitating effective comparative analysis during model training.

erated images are realistic and stylistically consistent with the test set, we use captions from the OOPS! dataset. The C_n and C_a captions are input into the Realistic Vision V4.0 model [42] to generate high-quality pedestrian images. It is worth noting that, before image generation, we filter the captions to ensure they describe human subjects. For each caption, we extract the subject and verify if it refers to a person, such as "people", "he", "man", *etc*. This process yields $6,739$ C_n and $6,979$ C_a captions. Additionally, for cases where both C_n and C_a are retained, we concatenate them to form a new C_{a+}. The resulting $6,669$ C_{a+} captions serve as prompts for generating pedestrian images with anomaly behaviors, enhancing diversity. Despite the ability of the Realistic Vision V4.0 model to produce photorealistic images, it occasionally generates images with unreasonable human structures. To mitigate this issue, we use OpenPose [7] for human key point detection and eliminate images with structural errors. After generating 75 images per caption and applying image filtering, we obtain 1,013,605 high-quality, diverse pedestrian training images, with an anomalous to normal behavior ratio of approximately 3:2.

Text Diversifying via Re-captioning. Similar to the caption generation process for test set image, we re-caption each synthetic image via Qwen2-VL [4]. The minor errors contained in the generated caption can be acceptable for the training set. So we directly use the generated caption as a matched text description for each training image.

Attribute Annotation. Aiming at pedestrian anomaly retrieval, we further enrich our dataset with annotations of actions, anomalies, and scenes. Given the cost of manually annotating large-scale data, we opt to leverage the multi-modal understanding capability of MLLM (Qwen2-VL [4]) to automatically obtain the action types, anomalous behaviors, and scenes for each image-text pair. Specially, for an anomaly image-text pair (I, T), we get its aforementioned attributes by querying the MLLM with carefully crafted instructions in *Suppl*. Note that the attribute recognition is not the focus of this paper, we provide these attributes to support future tasks like action or scene classification.

3.3. Dataset Analysis

Following the steps above, we have successfully created the Pedestrian Anomaly Behavior (PAB) dataset, a large-scale, richly annotated collection. Its properties and examples are depicted in Fig. 2. We compare PAB with other prominent Text-based Person Re-Identification datasets (*i.e.*, CUHK-PEDES [25], ICFG-PEDES [11], and RSTPReid [72]) and Video Anomaly Detection datasets (such as ShanghaiTech [33] and UBnormal [1]) in Table 1, focusing on image count, text annotation quantity and length, data sources, and annotation types. PAB features the following characteristics: **(1) A Large Number of Anomalous Behaviors:** Unlike general pedestrian datasets that focus on the viewpoint and occlusion of pedestrian images, PAB emphasizes providing a large number of images and textual descriptions of pedestrian anomaly. This complements ordinary Person Re-Identification tasks and presents new challenges. **(2) High-Fidelity Images**: Compared to pedestrian images from surveillance cameras, which often suffer from poor lighting, blurry textures, the images in PAB are of higher quality due to the generation method we adopted. The synthesized images are reasonable, realistic, and aesthetically pleasing. **(3) Specific Textual Descriptions**: Relative to existing cross-modal pedestrian datasets, the texts in PAB are longer and provide more detailed information, including the appearance, clothing, actions, and background context of individuals. For pedestrian anomaly retrieval, information about the appearance and actions of individuals is crucial. **(4) Diversity**: PAB encompasses a wide range of images that vary in terms of appearance, posture, viewpoint, background, and occlusion. Additionally, the text generation approach ensures that the image captions in PAB are sufficiently diverse, too. **(5) Large-Scale Image-Text Pairs**: PAB comprises over one million image-text pairs, supporting deep cross-modal models in learning better uni-modal features and more inter-modal associations from the data. **(6) Rich Annotations**: Each image-text pair in PAB is annotated with corresponding action, anomaly, and scene cat-
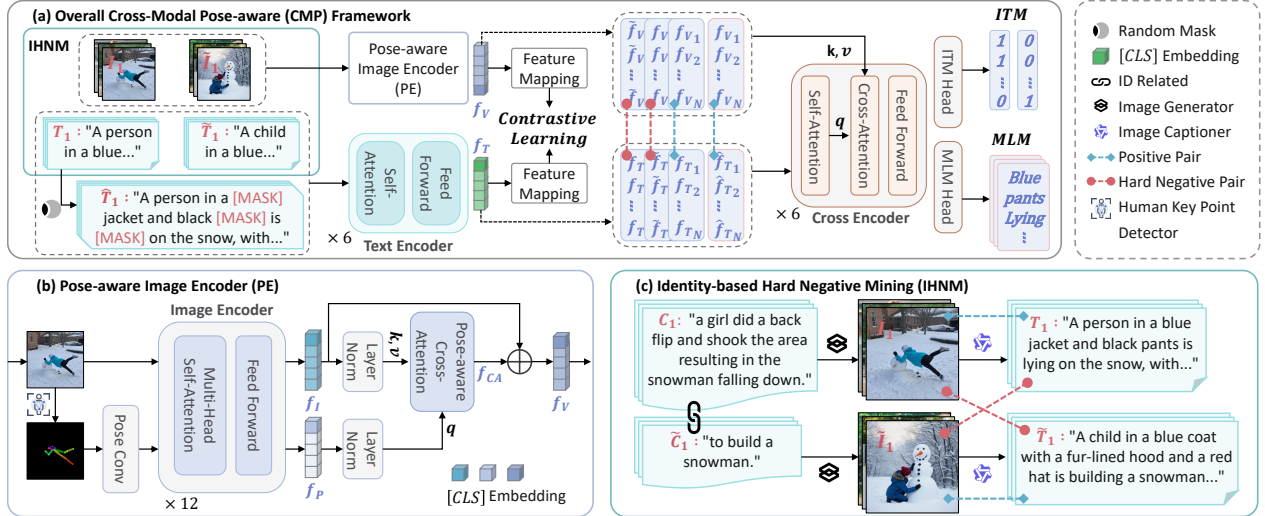
Figure 3. **(a) Overview of our Cross-Modal Pose-aware (CMP) framework**, composed of (b) a pose-aware image encoder, a text encoder, and a cross encoder. We apply (c) Identity-based Hard Negative Mining to form challenging negative pairs during training, followed by feature extraction, contrastive learning, and processing by the cross encoder with Image-Text Matching (ITM) and Mask Language Modeling (MLM) heads. The final step computes a cross-modal loss including ITM and MLM components.

egories, providing additional data information besides person appearance. **(7) Less Privacy Concerns**: Apart from a small amount of data in the test set sourced from existing public datasets, the vast majority of the data in PAB is synthesized, reducing ethical and legal issues.

4. Method

As shown in Fig. 3, we introduce Cross-Modal Pose-aware (CMP) framework, which includes a pose-aware image encoder, a text encoder, and a cross encoder. During training, we use our Identity-based Hard Negative Mining (IHNM) strategy to create challenging negative pairs. These pairs then undergo feature extraction through the respective encoders, followed by contrastive learning. The features are then processed by the cross encoder for multimodal encoding and pass through the Image-Text Matching (ITM) and Mask Language Modeling (MLM) heads. The final step involves calculating the IHNM-enhanced ITM loss and MLM loss. Hereinafter, we detail the three core components: the pose-aware image encoder, the identity-based hard negative mining strategy, and the cross-modal modeling module.

4.1. Pose-aware Image Encoder (PE)

Considering that normal and anomaly actions notably differ in human pose, we devise a pose-aware image encoder that explicitly incorporates human pose to enhance behavior comprehension. For a given image I, a human key point detector [7] is utilized to obtain the pose map P. Given P, we extract pose-aware feature via the Pose Conv module, and then feed the feature into the image encoder to obtain f_P, while I is directly input into the image encoder to extract image embedding f_I. The Pose Conv module comprises convolutional layers, batch normalization, and ReLU acti-

vation functions to align the input domain.

Image Encoder. Without loss of generality, we deploy Swin Transformer (Swin-B) [30] as the Image Encoder. The input image is initially divided into non-overlapping patches, which are then linearly embedded. These embedded patches are subsequently processed by transformer blocks, comprising Multi-Head Self-Attention and Feed Forward modules, to generate patch embeddings. Each patch embedding encapsulates the information of its corresponding patch. To aggregate the information from all patches, we compute the average of their features, referred to as the [CLS] embedding, and prepend it to the sequence.

Pose-aware Cross-Attention. Pose representation f_P, after being regularized by Layer Norm, is integrated into the image representation through a multi-head cross-attention module. The output (f_{CA}) of the cross-attention can be defined as: $f_{CA} = Softmax(\frac{qk^T}{\sqrt{d}})v$, where $q = W_q f_P, k = W_k f_I, v = W_v f_I$. q, k, v are the query, key, and values matrices of the attention operation respectively, d is the dimension of k, and W_q, W_k, W_v are the weights of the linear layers. Then, we combine the output of CA with the image embedding as the pose-aware representation $f_V = f_I + f_{CA}$.

4.2. Identity-based Hard Negative Mining (IHNM)

The key to solving pedestrian anomaly research is effectively establishing the relationship between pedestrian descriptions (with normal or anomaly action) and images. Ideally, for each text, the model should be able to search for an image that matches the appearance, actions, and background described, particularly the action description. Training such a cross-modal search model requires large-scale matching image-text pairs, negative pairs, and hard negative pairs, which are crucial for enabling the model to learn more

discriminative features. Based on the construction process of PAB, we propose an Identity-based hard negative pair mining (IHNM) method that provides corresponding hard negative samples for each image and text in the training set. The sampling strategy is illustrated in Figure 3 (c). Consider a training pair (I, T), where I is generated from anomaly/normal caption $C \in \{C_a, C_{a+}, C_n\}$, and T is its re-captioned text. As the normal/anomaly counterpart $\tilde{C}$ of C (both describing the same person) is known according to the OOPS! dataset, we can pick $\tilde{I}$ generated from $\tilde{C}$, and its re-captioned text $\tilde{T}$ can also be chosen accordingly. As shown in the Figure 3 (c), C and $\tilde{C}$ describe the anomalous and normal behaviors of the same pedestrian, respectively. Consequently, the generated images have similar pedestrian appearances and backgrounds but different actions. Intuitively, the image I and the text $\tilde{T}$ form an ID-based hard negative pair. The establishment of hard negative pairs with similar appearances and backgrounds but different actions is critical for the model to learn features that are more discriminative of actions.

4.3. Cross-modal Modeling

As shown in Figure 3, we optimize our proposed Cross-Modal Pose-aware model by three cross-modal modeling tasks, *i.e.*, Image-Text Matching (ITM), Contrastive Learning, and Masked Language Modeling (MLM). These tasks align texts and images integrated with the pose.

IHNM-Enhanced Image-Text Matching. Image-text matching determines whether an (image, text) pair is matched or not. Typically, well-aligned pairs are considered positive samples. Rather than randomly selecting trivial negatives, we employ our IHNM strategy to identify hard negative pairs. These pairs differ only in the action with its positive counterpart, requiring the model to discern subtle action differences, a capability crucial for the task of person anomaly search. Given an image-text pair, their multi-modal feature is encoded by feeding their representation into a cross encoder. Specifically, we use the last six layers of BERT as the cross encoder. The cross encoder takes the text embeddings as input and fuses the image embeddings in the cross-attention module at each layer. Similar to the multi-head cross-attention in PE, the text embeddings serve as queries (q), while the image embeddings act as keys (k) and values (v). The [CLS] embedding of features output by the cross encoder, is projected into 2-dimensional space via an ITM head (*i.e.*, an MLP), yielding the predicted image-text matching probability $\hat{p}(I, T)$. The Image-Text Matching loss is defined as:

$$
\begin{aligned}
\mathcal{L}_{\text{itm}} = -\mathbb{E}[& p(I, T) \log \hat{p}(I, T) \\
& + (1 - p(I, T)) \log(1 - \hat{p}(I, T))],
\end{aligned} \tag{1}
$$

where p denotes the ground-truth label. $p(I, T) = 1$ for positive pairs, 0 for negative ones.

Image-Text Contrastive Learning. This constrain targets to align image-text pair via contrastive learning [5, 22, 59]. Given the pose-aware image feature f_V, we split it to $\{f_{cls}, f_{pat}\}$, where cls and pat are the CLS embedding and patch embeddings, respectively. To obtain a comprehensive global image representation, we consider both the patch embedding and the CLS embedding: $f_v = FC([\text{AVG}(f_{pat}), f_{cls}])$, where $\text{AVG}(f_{pat})$ is the average feature of tokens in f_{pat}. Following previous practice [5, 59], we deploy the initial six transformer layers of BERT [10] for text encoding. The text T is tokenized and prefixed with a single [CLS] token before being input into the encoder. This process yields a text embedding f_T. Mirroring the image encoding, we apply the same procedure to f_T, combining text and class embeddings to obtain the global text representation f_t. The image-to-text similarity within the batch is defined as follows:

$$
S_{\text{I2T}} = \frac{\exp(s(f_v, f_t)/\tau)}{\sum_{j=1}^{N} \exp(s(f_v, f_t^j)/\tau)}, \tag{2}
$$

where $s(\cdot, \cdot)$ is the cosine similarity, τ is a temperature parameter. Similarly, the text-to-image similarity is S_{T2I}. Finally, the contrastive learning loss is presented below:

$$
\mathcal{L}_{\text{cl}} = -\frac{1}{2}\mathbb{E}[\log S_{\text{I2T}} + \log S_{\text{T2I}}]. \tag{3}
$$

Mask Language Modeling (MLM) predicts the masked words in the text based on the matched image. We randomly enable the masking strategy[*] with a probability of 25%. The person image I and the corresponding masked text $\hat{T}$ pair are then fed into the cross encoder, followed by an MLM head (an MLP with softmax) to predict the masked tokens. We minimize the cross-entropy loss:

$$
\mathcal{L}_{\text{mlm}} = -\mathbb{E}[p_{\text{mask}}(I, \hat{T}) log(\hat{p}_{\text{mask}}(I, \hat{T}))], \tag{4}
$$

where $\hat{p}_{\text{mask}}$ is the predicted likelihood of the masked token t in $\hat{T}$, p_{mask} is the ground-truth one-hot vector. The overall objective can be formulated as: $\mathcal{L} = \mathcal{L}_{\text{cl}} + \mathcal{L}_{\text{itm}} + \mathcal{L}_{\text{mlm}}$.

5. Experiment

Dataset and Evaluation Metrics. The constructed PAB serves as the evaluation benchmark. To evaluate the performance of text-based person anomaly search, we adopt recall rates R@K and mAP. Given a query text, all test images are ranked according to their matching probability with the query. If the image that perfectly matches the content of the text description (appearance, actions, background, *etc.*) is among the top k images, the search is considered successful. Mean Average Precision (mAP) refers to the average area under the precision-recall curve across all queries. Higher recall rates and mAP indicates better results.

---

[*]The masking strategy is: 10% of the tokens are replaced with random tokens, 10% remain unchanged, and 80% are replaced with [MASK].

Method	Normal		Wind		Rain		Snow		Rain + Snow		Dark		Dark + Wind		Dark + Rain		Dark + Snow		Over-exposure		Mean ↑	
	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP
Baseline	83.47	90.94	79.02	88.10	54.40	67.40	59.10	72.34	49.95	62.09	79.58	88.24	75.53	85.47	34.93	45.98	50.25	63.32	74.87	84.82	64.11	74.87
+ PE	84.13	91.20	78.92	87.93	55.01	67.31	61.78	74.11	50.76	63.28	78.41	87.72	75.43	85.43	39.08	49.77	50.10	63.06	75.89	85.51	64.95	75.53
+ IHNM	84.48	91.36	79.47	88.25	58.24	70.29	60.11	72.78	51.47	63.87	80.18	88.83	75.78	85.71	**40.75**	**50.88**	52.12	64.52	**76.64**	**85.99**	65.92	76.25
CMP (Ours)	**84.93**	**91.66**	**81.24**	**89.34**	**60.06**	**72.53**	**63.40**	**75.74**	**54.85**	**67.31**	**80.89**	**89.00**	**77.20**	**86.55**	39.03	50.58	**53.49**	**66.12**	76.09	85.56	**67.12**	**77.44**

Table 2. Robust text-based person anomaly retrieval results on PAB under multi-weather setting.

Method	#Data	R@1	R@5	R@10	mAP
MRA [60]	-	9.91	23.66	31.45	17.15
RaSa [5]	-	21.74	27.30	27.96	24.35
WoRA [46]	-	22.25	45.91	53.54	33.39
APTM [59]	-	22.90	45.80	52.38	33.56
CAMeL [62]	-	24.47	50.00	58.75	36.75
IRRA [22]	-	30.59	59.61	68.91	44.41
CLIP [39]	-	47.57	81.55	89.03	62.73
X-VLM [65]	-	71.94	97.78	98.99	83.96
MRA [60]	0.1M	70.53	94.69	97.47	81.59
APTM [59]	0.1M	72.14	95.30	97.17	82.78
CAMeL [62]	0.1M	74.30	96.79	98.84	84.20
WoRA [46]	0.1M	74.47	96.82	98.48	84.60
IRRA [22]	0.1M	76.39	97.62	99.14	86.33
CLIP [39]	0.1M	77.60	98.84	**99.75**	87.35
RaSa [5]	0.1M	80.79	<u>98.89</u>	99.65	89.20
X-VLM [65]	0.1M	81.95	98.84	99.19	89.86
CMP (Ours)	0.1M	<u>83.06</u>	<u>98.89</u>	99.49	<u>90.41</u>
CMP (Ours)	1M	**84.93**	**99.09**	**99.75**	**91.66**

Table 3. **Quantitative results of our proposed method and compared methods on the proposed dataset PAB.** The best result is indicated in **bold**, while the second best is <u>underlined</u>.

Implementation Details. We train the Cross-Modal Pose-aware (CMP) model for 30 epochs with mini-batch size of 22. We adopt AdamW [31] optimizer with a weight decay of 0.01. The learning rate linearly decreases from 1×10^{-4} to 1×10^{-5}. CMP has $230.8M$ trainable parameters, with $86.8M$, $66.4M$, and $59.1M$ parameters for the image, text, and cross encoders, respectively. The encoder weights are initialized by X-VLM [65].

Quantitative Results. In Table 3, we compare our method with a wide range of possible solutions, including two state-of-the-art vision-language pre-training models (CLIP [39] and X-VLM [65]) and six state-of-the-art text-based person search methods. Firstly, X-VLM [65] achieves the fairly-good zero-shot results, with 71.94% R@1, and 83.96% mAP, indicating that the fine-grained action understanding remains a significant challenge. After training 30 epochs on the same $0.1M$ image-text pairs from PAB, X-VLM [65] achieves 81.95% R@1, significantly higher than the other four methods. Therefore, we adopt the X-VLM [65] model to form a strong baseline. Compared to Baseline, our Cross-Modal Pose-aware (CMP) method employs Pose-aware Image Encoder (PE) and Identity-based Hard Negative Mining (IHNM) and achieves an R@1 of 83.06%, representing a $+1.11\%$ increase over the Baseline model. If all PAB data $(1M)$ is adopted to train CMP, the recall@1 is 84.93%, improving $+1.87\%$ compared to $0.1M$ training images. Similar trends are observed in other evaluation metrics.

Multi-weather Setting. Following MuSe-Net [48], we in-

Method	R@1	R@5	R@10	mAP
APTM [59]	27.86	40.41	46.77	22.61
IRRA [22]	40.28	57.24	65.98	33.53
CLIP [39]	51.60	68.31	76.43	43.05
X-VLM [65]	52.33	66.73	72.54	40.87
RaSa [5]	<u>54.12</u>	70.32	75.96	39.71
CMP	<u>54.12</u>	<u>71.07</u>	<u>77.90</u>	<u>43.13</u>
CMP (1M)	**55.23**	**71.67**	**77.99**	**44.35**

Table 4. Comparisons with existing methods in OOD setting. The unseen test set UCC is extracted from the UCF-Crime [45] dataset.

Evaluation setting	R@1	R@5	R@10	mAP
Identity Search (Traditional)	94.34	99.39	99.85	88.01
Behavior Search (Ours)	84.93	99.09	99.75	91.66

Table 5. Comparison of different evaluation settings with CMP.

Method	PE	IHNM	R@1	R@5	R@10	mAP
Baseline			83.47	99.04	**99.80**	90.94
M1	✓		84.13	**99.14**	99.65	91.20
M2		✓	84.48	98.94	99.60	91.36
CMP (Ours)	✓	✓	**84.93**	99.09	99.75	**91.66**

Table 6. Ablation studies on the key component of our proposed method, *i.e.*, Pose-aware Image Encoder (PE) and Identity-based Hard Negative Mining (IHNM).

troduce a multi-weather test setting to simulate the round-the-clock (24/7) smart city scenario. Specifically, we evaluate the CMP framework with and without the key component, *i.e.*, Pose-aware Image Encoder (PE) and Identity-based Hard Negative Mining (IHNM), under 10 distinct environmental conditions, *e.g.*, wind/rain/snow. As shown in Table 2, CMP proves pivotal for robustness, showing consistent gains across all scenarios.

OOD Setting. To evaluate the scalability of the CMP model, we extract a new real-world test set from UCF-Crime [45] for an Out-of-Distribution (OOD) test, and we call it UCC. Specifically, we select keyframes from 13 types of abnormal and normal videos in UCF-Crime and leverage Qwen2-VL [4] to generate text descriptions for these video frames. It results $5,320$ image-text pairs as the OOD test set. As shown in Table 4, our model CMP achieves superior performance, 54.12% R@1 and 43.13% mAP, compared to other state-of-the-art methods (also trained on $0.1M$ PAB data). After training with complete $1M$ PAB training data, CMP further improves by 1.11% in R@1. These experimental results show that our model exhibits competitive adaptability when facing unseen datasets.

Behavior Search vs. Identity Search. Text-based person anomaly search poses significantly greater challenges than traditional identity-based person search, as it requires
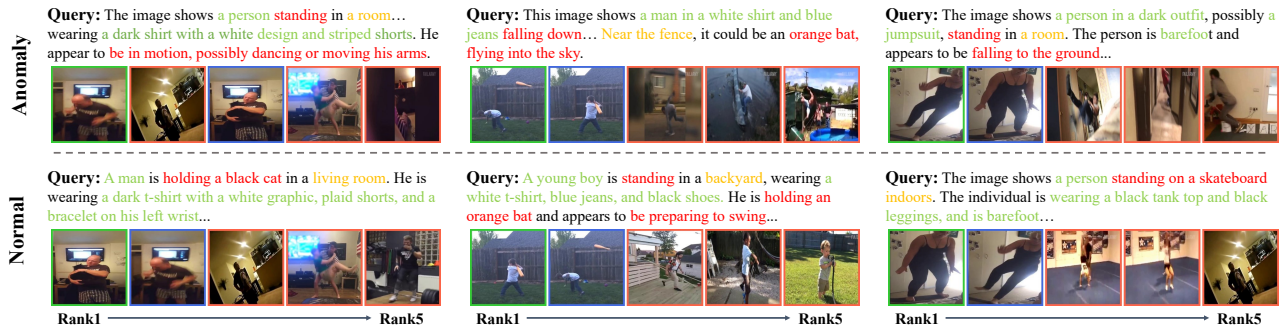
Figure 4. Qualitative Results. Top-5 anomaly search results for text queries: anomaly actions (top) and normal actions (bottom). Green rectangles indicate correct matches, red for mismatches, and blue for ID matches with behavior mismatches. Query parts (appearance, action, background) are highlighted in green, red, and orange, respectively.

a fine-grained understanding of both appearance and behavior. As shown in Table 5, while traditional methods treat all images with the same ID as correct matches, our anomaly search task demands precise localization of specific behaviors. This distinction is evident in our experimental results: our model achieves 94.34% R@1 on traditional text-based person search, outperforming its performance on anomaly search by 9.41%. The performance gap stems from the fundamental difference between the two tasks: identity search primarily relies on appearance, whereas anomaly retrieval necessitates distinguishing subtle behavioral patterns. Such capability is crucial for real-world applications, including security event tracing, emergency response, and other scenarios where requires precise behavior localization.

Qualitative Results. We qualitatively evaluate our method on the task of Text-based Person Anomaly Search. Figure 4 shows six fine-grained pedestrian behavior retrieval results via the trained Cross-Modal Pose-aware (CMP) model, with queries for anomalous behavior (top) and normal behavior (bottom). For each query, we display the top five retrieved images. Correct retrieval results are marked with green boxes. Blue and red boxes indicate incorrect matches. Note that blue boxes denote images that belong to the same ID as the query but do not match the action. To further illustrate the retrieval effects, we highlight the parts of the text queries that describe appearance in green, action in red, and the background in orange. It can be observed that, in addition to appearance and background information, our model can effectively distinguish fine-grained action information. Even the incorrectly matched images displayed in Figure 4 still show reasonable relevance to the query sentences.

Effect of the Key Components. We conduct ablation studies on the key components of the Cross-Modal Pose-aware (CMP) model. As shown in Table 6, for fairness, all variants are trained for 30 epochs on $1M$ image-text pairs. First, we evaluate the effectiveness of Pose-aware Image Encoder (PE) by comparing the Baseline and M1 models. It indicates that Pose-aware Image Encoder (PE) is a critical component of CMP. Furthermore, M2 incorporating Identity-based Hard Negative Mining (IHNM) obtains 1.01% R@1
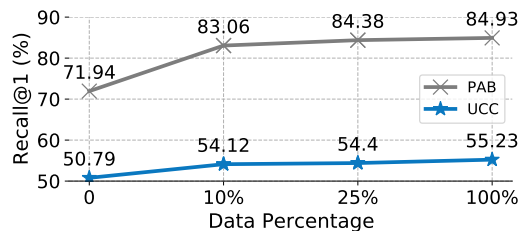

Figure 5. Ablation studies on synthetic training data scale. We gradually increase PAB training data from 0% to 100% .

improvement compared to the Baseline. The results show that adding IHNM improves the ability in discriminating the fine-grained behavior. Our CMP method equipped with both PE and IHNM arrives at the best performance.

Effect of Training Data. We further study the impact of training data scale on model performance in Fig. 5. Considering the hard negative sampling, our method with 10% training data has already achieves competitive recall rate.

6. Conclusion

We propose Text-based Person Anomaly Search, a new task that overcomes the limitation of traditional retrieval in anomaly identification. To support this, we introduce the large-scale Pedestrian Anomaly Behavior (PAB) benchmark, combining synthetic training data with real-world test frames. Our Cross-Modal Pose-aware (CMP) framework leverages pose patterns and hard negative mining to distinguish normal/anomalous actions. Extensive experiments on PAB and UCC confirm the effectiveness of CMP in retrieving both anomalies and normal behaviors in real-world scenarios and multi-weather settings.

References

[1] Andra Acsintoae, Andrei Florescu, Mariana-Iuliana Georgescu, Tudor Mare, Paul Sumedrea, Radu Tudor Ionescu, Fahad Shahbaz Khan, and Mubarak Shah. Ubnormal: New benchmark for supervised open-set video anomaly detection. In *CVPR*, pages 20143–20153, 2022. 2, 3, 4, 13

[2] Amit Adam, Ehud Rivlin, Ilan Shimshoni, and Daviv Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE transactions on pattern analysis and machine intelligence*, 30(3):555–560, 2008. 2, 3, 13

[3] Jon Almazan, Bojana Gajic, Naila Murray, and Diane Larlus. Re-id done right: towards good practices for person re-identification. *arXiv:1801.05339*, 2018. 2

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv:2308.12966*, 2023. 3, 4, 7, 12

[5] Yang Bai, Min Cao, Daming Gao, Ziqiang Cao, Chen Chen, Zhenfeng Fan, Liqiang Nie, and Min Zhang. Rasa: relation and sensitivity aware representation learning for text-based person search. In *IJCAI*, pages 555–563, 2023. 3, 6, 7

[6] Congqi Cao, Yue Lu, and Yanning Zhang. Context recovery and knowledge retrieval: A novel two-stream framework for video anomaly detection. *IEEE Transactions on Image Processing*, 33:1810–1825, 2024. 3

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, pages 7291–7299, 2017. 3, 4, 5

[8] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv:2306.15195*, 2023. 3

[9] Yuhao Chen, Guoqing Zhang, Yujiang Lu, Zhenxing Wang, and Yuhui Zheng. Tipcb: A simple but effective part-based convolutional baseline for text-based person search. *Neurocomputing*, 494:171–181, 2022. 3

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 6

[11] Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv:2107.12666*, 2021. 1, 2, 3, 4

[12] Zhongjie Duan, Chengyu Wang, Cen Chen, Wenmeng Zhou, Jun Huang, and Weining Qian. Match4match: Enhancing text-video retrieval by maximum flow with minimum cost. In *WWW*, pages 3257–3267, 2023. 3

[13] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *CVPR*, pages 919–929, 2020. 3

[14] Xinyang Feng, Dongjin Song, Yuncong Chen, Zhengzhang Chen, Jingchao Ni, and Haifeng Chen. Convolutional transformer based dual discriminator generative adversarial networks for video anomaly detection. In *ACM MM*, pages 5546–5554, 2021. 3

[15] Alessandro Flaborea, Luca Collorone, Guido Maria D'Amely Di Melendugno, Stefano D'Arrigo, Bardh Prenkaj, and Fabio Galasso. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In *ICCV*, pages 10318–10329, 2023. 3

[16] Chenyang Gao, Guanyu Cai, Xinyang Jiang, Feng Zheng, Jun Zhang, Yifei Gong, Pai Peng, Xiaowei Guo, and Xing Sun. Contextual non-local alignment over full-scale representation for text-based person search. *arXiv:2101.03036*, 2021. 3

[17] Ping Guo, Yue Hu, Yanan Cao, Yubing Ren, Yunpeng Li, and Heyan Huang. Query in your tongue: Reinforce large language models with retrievers for cross-lingual search generative experience. In *WWW*, pages 1529–1538, 2024. 3

[18] Kai Han, Jianyuan Guo, Chao Zhang, and Mingjian Zhu. Attribute-aware attention model for fine-grained representation learning. In *ACM MM*, pages 2040–2048, 2018. 2

[19] Xiao Han, Sen He, Li Zhang, and Tao Xiang. Text-based person search with limited data. *arXiv:2110.10807*, 2021. 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[21] Or Hirschorn and Shai Avidan. Normalizing flows for human pose anomaly detection. In *ICCV*, pages 13545–13554, 2023. 3

[22] Ding Jiang and Mang Ye. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *CVPR*, pages 2787–2797, 2023. 3, 6, 7

[23] Xintong Jiang, Yaxiong Wang, Yujiao Wu, Bingwen Hu, and Xueming Qian. Cala: Complementary association learning for augmenting composed image retrieval. *SIGIR*, 2024. 3

[24] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3

[25] 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. 1, 2, 3, 4

[26] Shiping Li, Min Cao, and Min Zhang. Learning semantic-aligned feature representation for text-based person search. In *ICASSP*, pages 2724–2728. IEEE, 2022. 3

[27] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2013. 2, 3, 13

[28] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang. Improving person re-identification by attribute and identity learning. *Pattern recognition*, 95:151–161, 2019. 2

[29] Hefei Ling, Ziyang Wang, Ping Li, Yuxuan Shi, Jiazhong Chen, and Fuhao Zou. Improving person re-identification by multi-task learning. *Neurocomputing*, 347:109–118, 2019. 2

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 5

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 7

[32] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *ICCV*, pages 2720–2727, 2013. 2, 3, 13

[33] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *ICCV*, pages 341–349, 2017. 2, 3, 4, 13

[34] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. Aggregating deep pyramidal representations for person re-identification. In *CVPR workshops*, pages 0–0, 2019. 2

[35] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, pages 935–942. IEEE, 2009. 2, 13

[36] Kai Niu, Yan Huang, Wanli Ouyang, and Liang Wang. Improving description-based person re-identification by multi-granularity image-text alignments. *IEEE Transactions on Image Processing*, 29:5542–5556, 2020. 3

[37] Jicheol Park, Dongwon Kim, Boseung Jeong, and Suha Kwak. Plot: Text-based person search with part slot attention for corresponding part discovery. In *ECCV*, pages 474–490. Springer, 2024. 3

[38] Xueming Qian, Dan Lu, Yaxiong Wang, Li Zhu, Yuan Yan Tang, and Meng Wang. Image re-ranking based on topic diversity. *IEEE Trans. Image Process.*, 26(8):3734–3747, 2017. 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 3, 7

[40] Bharathkumar Ramachandra and Michael Jones. Street scene: A new dataset and evaluation protocol for video anomaly detection. In *WACV*, pages 2569–2578, 2020. 2, 13

[41] Tal Reiss and Yedid Hoshen. Attribute-based representations for accurate and interpretable video anomaly detection. *arXiv:2212.00789*, 2022. 3

[42] SG_161222. Realvisxl_v4.0, 2024. https://huggingface.co/SG161222/RealVisXL_V4.0. 4

[43] Zhiyin Shao, Xinyu Zhang, Meng Fang, Zhifeng Lin, Jian Wang, and Changxing Ding. Learning granularity-unified representations for text-to-image person re-identification. In *ACM MM*, pages 5566–5574, 2022. 3

[44] Xiujun Shu, Wei Wen, Haoqian Wu, Keyu Chen, Yiran Song, Ruizhi Qiao, Bo Ren, and Xiao Wang. See finer, see more: Implicit modality alignment for text-based person retrieval. In *ECCV workshop*, 2023. 3

[45] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *CVPR*, pages 6479–6488, 2018. 2, 3, 7, 13

[46] Jintao Sun, Hao Fei, Gangyi Ding, and Zhedong Zheng. From data deluge to data curation: A filtering-wora paradigm for efficient text-based person search. In *WWW*, pages 2341–2351, 2025. 7

[47] Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. Harnessing the power of mllms for transferable text-to-image person reid. In *CVPR*, pages 17127–17137, 2024. 3

[48] Tingyu Wang, Zhedong Zheng, Yaoqi Sun, Chenggang Yan, Yi Yang, and Tat-Seng Chua. Multiple-environment self-adaptive network for aerial-view geo-localization. *Pattern Recognition*, 152:110363, 2024. 7

[49] Xiaodong Wang, Hongmin Hu, Fei Yan, Junwen Lu, Zhiqiang Zeng, Weidong Hong, and Zhedong Zheng. Uniad: Integrating geometric and semantic cues for unified anomaly detection. In *ACM Multimedia*, 2025. 2

[50] Yaxiong Wang, Hao Yang, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. Position focused attention network for image-text matching. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 3792–3798, 2019. 3

[51] Yaxiong Wang, Hao Yang, Xiuxiu Bai, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. PFAN++: bi-directional image-text retrieval with position focused attention network. *IEEE Trans. Multim.*, 23:3362–3376, 2021. 3

[52] Yaxiong Wang, Lianwei Wu, Lechao Cheng, Zhun Zhong, Yujiao Wu, and Meng Wang. Beyond general alignment: Fine-grained entity-centric image-text matching with multi-modal attentive experts. *SIGIR*, 2025. 3

[53] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Caibc: Capturing all-round information beyond color for text-based person retrieval. In *ACM MM*, pages 5314–5322, 2022. 3

[54] Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *ACM MM*, pages 1984–1992, 2022. 3

[55] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv:1901.11196*, 2019. 13

[56] Peng Wu, Jing Liu, Xiangteng He, Yuxin Peng, Peng Wang, and Yanning Zhang. Toward video anomaly retrieval from video anomaly detection: New benchmarks and model. *IEEE Transactions on Image Processing*, 33:2213–2225, 2024. 3

[57] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. Clip-driven fine-grained text-image person re-identification. *arXiv:2210.10276*, 2022. 3

[58] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. When urban region profiling meets large language models. In *WWW*, 2024. 3

[59] Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *ACM MM*, pages 4492–4501, 2023. 3, 6, 7

[60] Shuyu Yang, Yaxiong Wang, Yongrui Li, Li Zhu, and Zhedong Zheng. Minimizing the pretraining gap: Domain-aligned text-based person retrieval. *arXiv preprint arXiv:2507.10195*, 2025. 7

[61] Linli Yao, Weijing Chen, and Qin Jin. Capenrich: Enriching caption semantics for web images via cross-modal pretrained knowledge. In *WWW*, pages 2392–2401, 2023. 3

[62] Hang Yu, Jiahao Wen, and Zhedong Zheng. Camel: Cross-modality adaptive meta-learning for text-based person retrieval. *IEEE Transactions on Information Forensics and Security*, 2025. 7

[63] Tongtong Yuan, Xuange Zhang, Kun Liu, Bo Liu, Chen Chen, Jian Jin, and Zhenzhen Jiao. Towards surveillance video-and-language understanding: New dataset baselines and challenges. In *CVPR*, pages 22052–22061, 2024. 2, 3, 13

[64] M Zaigham Zaheer, Arif Mahmood, M Haris Khan, Mattia Segu, Fisher Yu, and Seung-Ik Lee. Generative cooperative learning for unsupervised video anomaly detection. In *CVPR*, pages 14744–14754, 2022. 3

[65] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *ICML*, 2022. 7

[66] Guoshuai Zhao, Chaofeng Zhang, Heng Shang, Yaxiong Wang, Li Zhu, and Xueming Qian. Generative label fused network for image-text matching. *Knowl. Based Syst.*, 263: 110280, 2023. 2

[67] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015. 2

[68] Zhedong Zheng and Liang Zheng. 2. object re-identification: Problems, algorithms and responsible research practice. *The Boundaries of Data*, page 21, 2024. 1

[69] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, pages 3754–3762, 2017. 2

[70] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *CVPR*, pages 2138–2147, 2019. 2

[71] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(2):1–23, 2020. 3

[72] Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *ACM MM*, pages 209–217, 2021. 1, 2, 4

Beyond Walking: A Large-Scale Image-Text Benchmark
for Text-based Person Anomaly Search
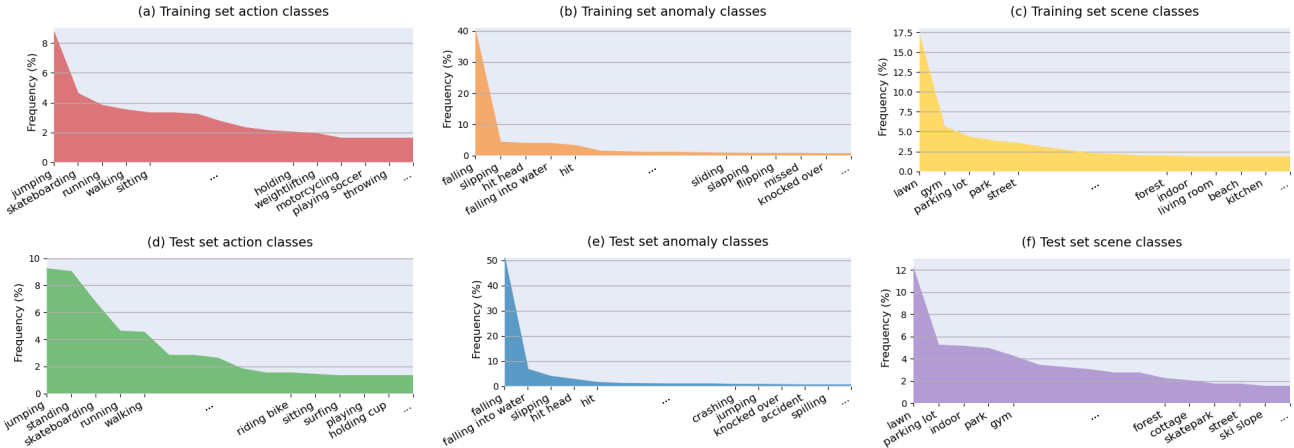
Supplementary Material

Figure 6. **Dataset Statistics**. An overview of the attribute annotations, including the distribution of categories across the training and test sets. Specifically, it covers normal action categories (a, d), anomaly categories (b, e), and scene categories (c, f). Due to the natural long-tail distribution of the data and the space limitation, we present the top 15 most common classes for each category to ensure clarity. (Best viewed when zooming in.)

Appendix

A. More Benchmark Details

A.1. Attribute Annotation Details.

During the attribute annotation process, we utilize the widely-used Qwen2-VL [4] to annotate each normal image-text pair with an action type and scene category, while each anomaly image-text pair is annotated with an anomalous behavior class and scene category. For a given image-text pair (I, T), if (I, T) is a training pair, I is generated from anomaly/normal caption $C \in \{C_a, C_{a+}, C_n\}$, and T is its re-captioned text. If (I, T) is a test pair, $C \in \{C_a, C_n\}$ is the caption of corresponding source video and T is the re-captioned text for I. We leverage I, C to design instructions and query the MLLM for attribute. The specific **Instructions** are as follows:

- **Instruction for Anomaly Behavior Class:** "Below is the image caption of the image. In the image, someone fails to do something. Based on the caption and image, summarize the failure of the characters in the image using a single word or phrase, such as falling, losing balance, slipping, falling to the ground, falling into water, losing control, having accident, flipping, jumping, hitting head, *etc*. Image caption: C."
- **Instruction for Action Type:** "Below is the image caption of the image. Based on the caption and image, summarize the behavior and action categories of the charac-

ters in the image using a single word or phrase, such as motorcycling, driving car, somersaulting, riding scooter, catching fish, staring at someone, dyeing eyebrows, trimming beard, peeling potatoes, square dancing, *etc*. Image caption: C."

- **Instruction for Scene Category:** "Below is the image caption of the image. Based on the caption and image, summarize the scene or background of the characters in the image using a single word or phrase, such as playground, parking lot, ski slope, highway, lawn, outdoor church, cottage, indoor flea market, fabric store, hotel, *etc*. Image caption: C."

A.2. Attribute Statistics.

Based on the three types of instructions, we automatically obtain action, anomaly, and scene attributes. As shown in Figure 6, we present the distribution of the top 15 most common classes for each attribute in both the training and test sets. The attribute distributions in both sets are similar and naturally exhibit a long-tail distribution For action types, the top five in the training set are jumping, skateboarding, running, walking, and sitting, while the top five in the test set are jumping, standing, skateboarding, running, and walking (Figure 6 (a) and (d)). The most frequent anomalous behavior is falling, occurring with approximately 40% frequency in the training set (Figure 6 (b)) and 50% in the test set (Figure 6 (e)). The scene distribution is primarily concentrated

Datasets	Modality	#Frames	#Scenes	#Anomaly Types	Anomaly:Normal	#Avg Words	Open Set	Data Source
UCSD Ped2 [27]	Video	4,560	1	5 Classes	1:2	-	✓	Collection
UMN [35]	Video	7,741	3	1 Classes	1:4	-	✓	Collection
UCSD Ped1 [27]	Video	14,000	1	5 Classes	1:2	-	✓	Collection
CUHK Avenue [32]	Video	30,652	1	5 Classes	1:7	-	✓	Collection
Subway Exit [2]	Video	64,901	1	3 Classes	1:13	-	✓	Collection
Subway Entrance [2]	Video	144,250	1	5 Classes	1:11	-	✓	Collection
Street Scene [40]	Video	203,257	1	17 Classes	1:4	-	✓	Collection
UBnormal [1]	Video	236,902	29	22 Classes	2:3	-	✓	Synthesis
ShanghaiTech [33]	Video	317,398	13	11 Classes	1:18	-	✓	Collection
UCF-Crime [45]	Video	13,741,393	Unlimited	13 Classes	≪1:1	-	×	Collection
UCA [63]	Video, Text	13,741,393	Unlimited	13 Classes	≪1:1	20.2	×	Collection
PAB (Ours)	Image, Text	1,015,583	**480**	**1600** Classes	**3:2**	50.3	✓	Synthesis & Collection

Table 7. Comparison of the statistics of our PAB and other Video Anomaly Detection (VAD) datasets. The statistics of previous datasets have been recorded in [1].

on the lawn, gym, and parking lot in both subsets, as shown in Figure 6 (c) and (f).

A.3. Comparisons with More Video Anomaly Detection Datasets.

In Table 7, we compare our proposed PAB dataset with the most utilized Video Anomaly Detection (VAD) datasets. Eight metrics are reported: modality, number of frames/images, scenes, anomaly types, the proportion of anomaly versus normal, average number of words per sentence, open-set characteristics, and data source. Compared to other video datasets, PAB is distinguished as an image-text pair dataset and features a higher number of anomaly types from a broader range of event scenes. For all datasets, the "Anomaly:Normal" ratio represents the proportion of anomaly video frames/images to normal video frames/images. While most VAD datasets are annotated solely with normal/abnormal labels or abnormal category labels, PAB provides detailed annotations including appearance descriptions, actions, and scene information. Most video datasets maintain open-set characteristics for anomaly detection. To ensure consistent open-set characteristics, we provide a real-world Out-of-Distribution (OOD) test set for PAB sourced from UCF-Crime [45]. Notably, UBnormal [1] is also a synthetic dataset, but unlike PAB, both its training and test sets consist entirely of synthesized data.

A.4. Visualizations.

In Figure 7, we present additional example image-text pairs from our proposed dataset, PAB. The figure includes 12 synthetic training image-text pairs (top) and 12 real-world test image-text pairs (bottom). These pairs are divided into two categories: one depicts anomalous behaviors, while the other illustrates normal actions. Each image-text pair is meticulously annotated with specific scene and action (or anomaly) classifications to facilitate further precise learning and evaluation. It is worth noting that while the training set sometimes contains some noise in the generated captions, the test set captions have been professionally refined to ensure high-quality annotations. This provides a reliable benchmark for assessing model performance.

B. Experiment Details and Further Experiments

B.1. Training Details.

We train the Cross-Modal Pose-aware (CMP) model using PyTorch on four NVIDIA GeForce RTX 3090 GPUs. The first 500 training iterations serve as a warm-up phase. Each image input is resized to 224×224 pixels, and the maximum text token length is set to 56. For image augmentation, we apply techniques such as random horizontal flipping and random erasing. For text augmentation, we employ EDA [55]. Training for 30 epochs on the full training set takes approximately 4 days and 4 hours.

B.2. Inference Details.

During inference, we first obtain the embeddings of all query texts and candidate images (integrated with pose) from the test set, then compute the text-to-image similarity. For each query, we select the top 128 images with the highest similarity scores. These images are then re-ranked based on the matching probabilities predicted by the cross-modal encoder and the MLP head. The final ranking results constitute the search outcomes of the model.

B.3. More Qualitative Result Examples.

We present 12 additional text-based person anomaly search qualitative results of our method in Figure 8. For each query (anomalous or normal), we display the top five retrieved images. True retrieval results are marked with green boxes,

Figure 7. **Dataset Examples**. 12 training (synthetic) image-text pairs from the PAB dataset are at the top, while 12 test (real-world) image-text pairs are at the bottom. Half of the examples depict anomaly behaviors, while the other half show corresponding normal actions. Each pair is annotated with scene and action (or anomaly) classes. Minor errors may be present in the generated captions of the training set, whereas the captions in the test set have been refined by professionals. (Best viewed on a computer screen with zoom.)

while blue and red boxes indicate incorrect matches. Blue boxes denote images that belong to the same ID as the query but do not match the action. We highlight the parts of the text queries that describe appearance in green, action in red, and the background in orange. In addition to appearance and background information, our model can effectively distinguish fine-grained action information. Even the incorrectly matched images displayed in Figure 8 still show some relevance to the query sentences.

Anomaly

Query: A child is standing on a concrete surface, wearing a patterned shirt and shorts. The child is holding a toolbox that is open on the ground, containing various items...

Query: The image shows a person riding a blue bicycle on a grassy field. He is about to fall down. The person is wearing a light blue hoodie and light-colored pants...

Query: The image shows a person standing on a white platform in a living room. She is waving her hands and feet, looking like she is about to fall. The individual is wearing...

Normal

Query: The image shows a young child walking on a concrete surface. The child is wearing a long-sleeved shirt and shorts with a colorful pattern...

Query: The image shows a person riding a bicycle on a grassy hill. The rider is wearing a blue helmet, a gray jacket, and dark pants. The background features a mix of green...

Query: The image shows a person standing on a white platform in a living room... black shorts, and she has one leg raised, with her arm extended upwards...

Anomaly

Query: A young man is bending over, and the basketball hoop above him is falling, which could hit his head in a backyard. He is wearing a blue shirt, black shorts, green...

Query: The image shows a person performing a flip in a grassy area... wearing a black outfit and appears to be in mid-air, with his legs extended and arms raised...

Query: The image shows two children playing on a trampoline in a garden. One child... is standing on the trampoline, while the other... is lying on the trampoline...

Normal

Query: A young person is playing soccer in a backyard. He is wearing a blue and black jacket, black pants, and white socks. The yard has a wooden fence, a basketball hoop...

Query: A person is balancing on a slackline outdoors, wearing a black hoodie, black shorts, and sneakers. The background features greenery and trees.

Query: The image shows two children standing on a trampoline in a garden. The girl, wearing a pink top and purple shorts, is reaching out to the boy, who is dressed...

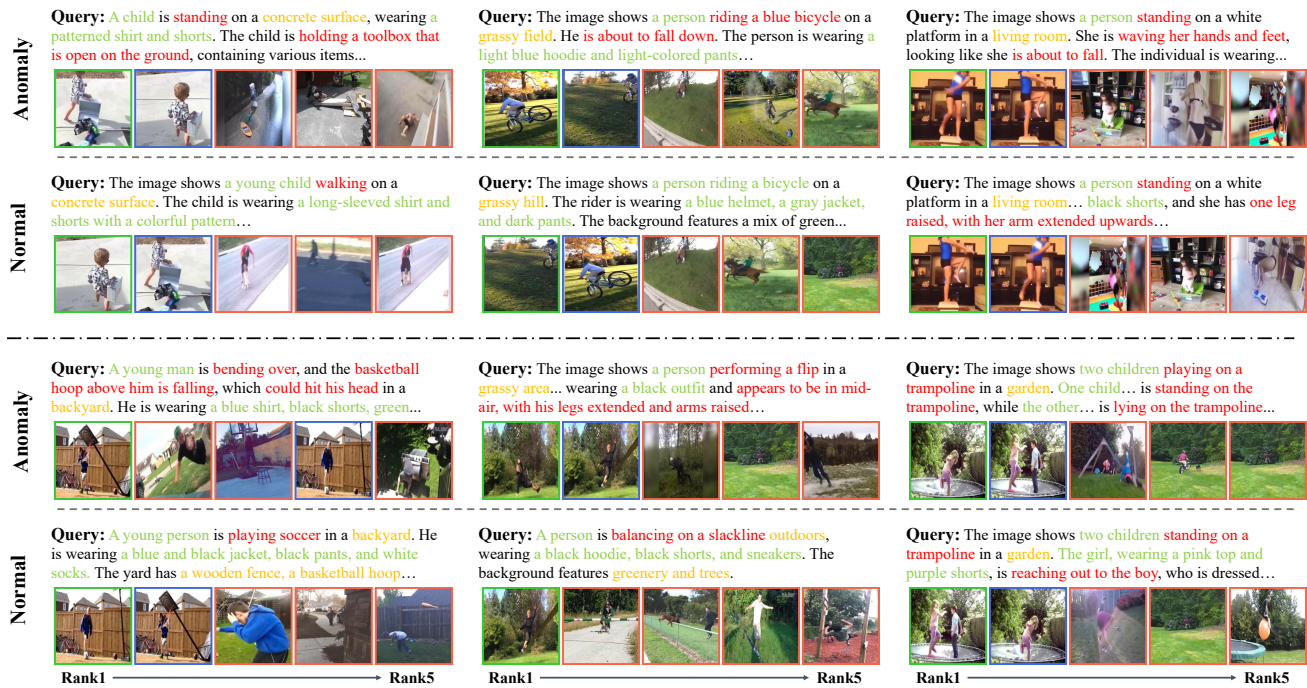Rank1 ⟶ Rank5   Rank1 ⟶ Rank5   Rank1 ⟶ Rank5

Figure 8. **More Qualitative Results.** 12 examples of top-5 person anomaly search results with text queries for anomaly actions and normal actions. Matched images are marked by green boxes, mismatched images are marked in red, and blue boxes indicate cases where the ID matches but the behavior does not. The parts of the queries that describe appearance, action, and background are highlighted in green, red, and orange. It is best viewed on a computer screen with Zoom.