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

Supplementary Material

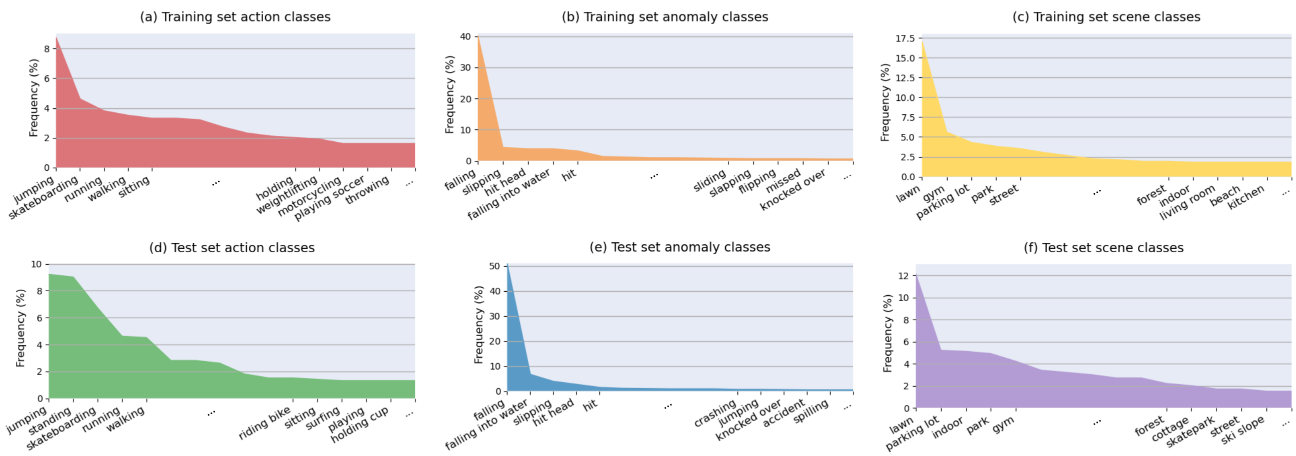

Figure 6. **Dataset Statistics**. An overview of the attribute annotations, including the distribution of categories across the training and test sets. Specifically, it covers normal action categories (a, d), anomaly categories (b, e), and scene categories (c, f). Due to the natural long-tail distribution 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 [3] 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 characters 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 [4]	Video	4,560	1	5 Classes	1:2	-	✓	Collection
UMN [7]	Video	7,741	3	1 Classes	1:4	-	✓	Collection
UCSD Ped1 [4]	Video	14,000	1	5 Classes	1:2	-	✓	Collection
CUHK Avenue [5]	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 [8]	Video	203,257	1	17 Classes	1:4	-	✓	Collection
UBnormal [1]	Video	236,902	29	22 Classes	2:3	-	✓	Synthesis
ShanghaiTech [6]	Video	317,398	13	11 Classes	1:18	-	✓	Collection
UCF-Crime [9]	Video	13,741,393	Unlimited	13 Classes	$\ll$1:1	-	×	Collection
UCA [11]	Video, Text	13,741,393	Unlimited	13 Classes	$\ll$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 [9]. 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 [10]. 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,

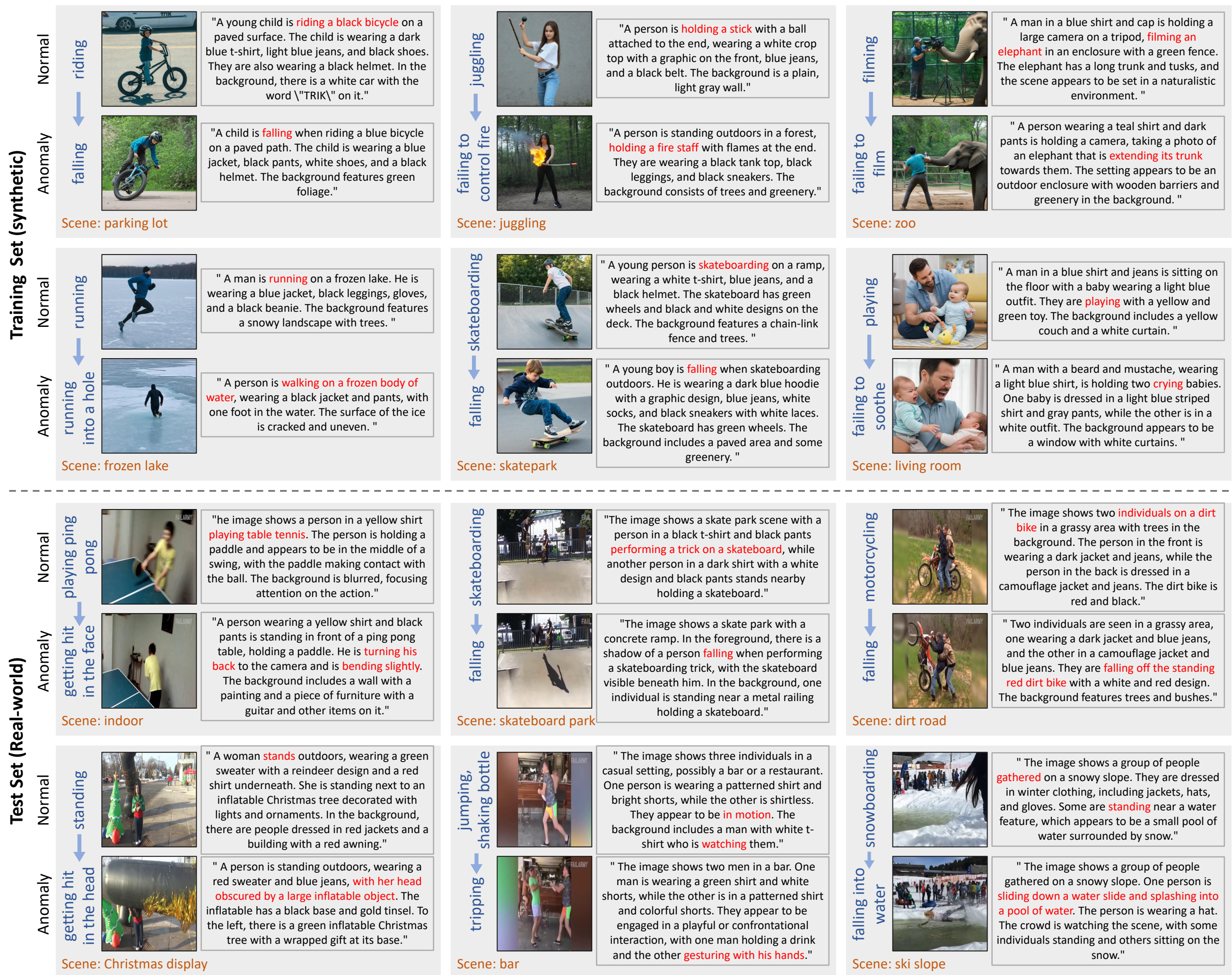

Figure 7. **Dataset Examples**. 12 training (synthetic) image-text pairs from the PAB dataset are at the top, while 12 test (real-world) image-text pairs are at the bottom. Half of the examples depict anomaly behaviors, while the other half show corresponding normal actions. Each pair is annotated with 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.

References

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

[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] 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. 1

[4] 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

[5] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detec-

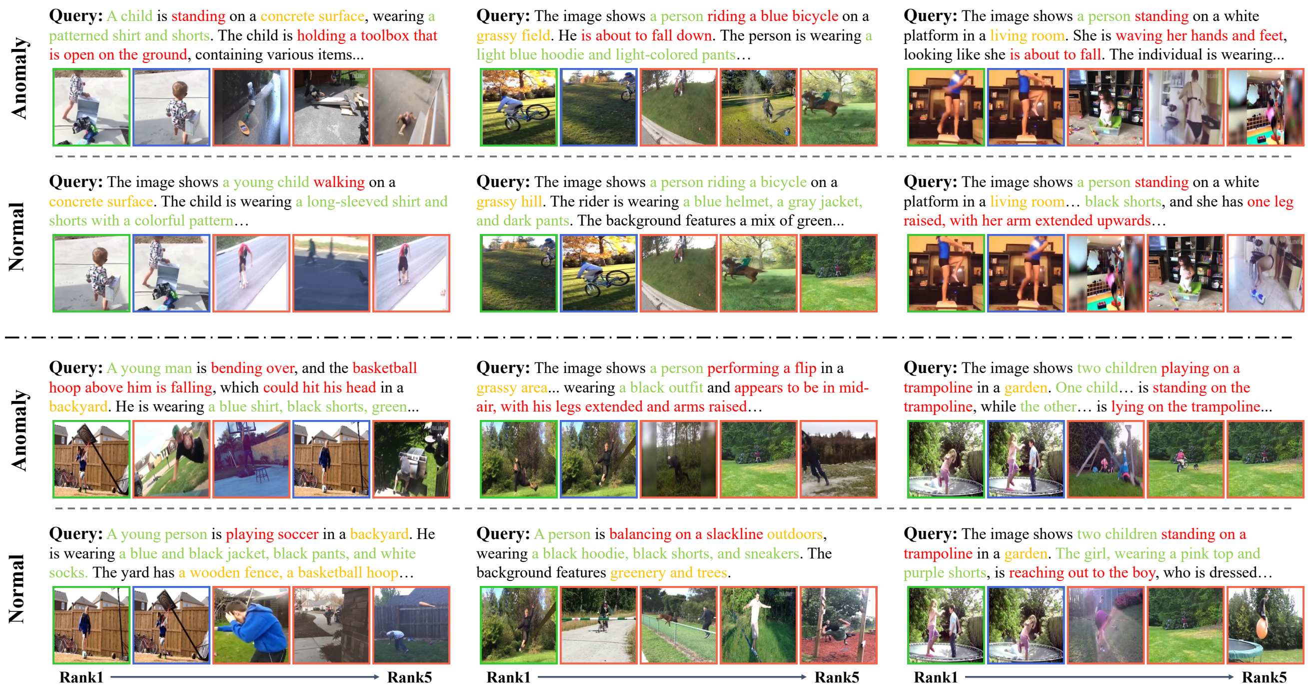

Figure 8. **More Qualitative Results.** 12 examples of top-5 person anomaly search results with text queries for anomaly actions and normal actions. Matched images are marked by green boxes, mismatched images are marked in red, and blue boxes indicate cases where the ID matches but the behavior does 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.

tion at 150 fps in matlab. In *ICCV*, pages 2720–2727, 2013. 2

[6] 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

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

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

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

[10] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv:1901.11196*, 2019. 2

[11] 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