# Bridging the Pose-Semantic Gap: A Cascade Framework for Text-Based Person Anomaly Search

**Zequn Xie**[1,†], **Guijin Luo**[1,†], **Chuxin Wang**[1], **Sihang Cai**[1], **Tao Jin**[1], **Zhou Zhao**[1], **Yixuan Tang**[2*]
[1] Zhejiang University
[2] National University of Singapore
Correspondence: zqxie@zju.edu.cn, yixuan@comp.nus.edu.sg

## Abstract

Text-based person anomaly search retrieves specific behavioral events from surveillance archives using natural-language queries. Although recent pose-aware methods align geometric structures well, they face a fundamental *Pose-Semantic Gap*: semantically different actions can share similar skeletal geometries. While Multimodal Large Language Models (MLLMs) can reduce this ambiguity, using them for large-scale retrieval is computationally prohibitive. We propose the *Structure-Semantic Decoupled Cascade (SSDC)* framework, which decouples retrieval into two stages: (1) *Structure-Aware Coarse Retrieval*, where a lightweight model quickly filters candidates by skeletal similarity; and (2) Detective Squad Interaction, a multi-agent semantic verification module. The squad consists of a *Detective* for fast binary filtering, an *Analyst* for evidence extraction, and a *Writer* for semantic synthesis. Finally, we re-rank candidates by fusing the synthesized captions with structural priors. Experiments on the PAB benchmark show that SSDC achieves state-of-the-art performance by balancing efficiency and semantic reasoning.Our code is available at: https://github.com/GridNexus/SSDC.

## 1 Introduction

Text-based person search (Li et al., 2017; Zheng and Zheng, 2024) has emerged as a critical technology in intelligent surveillance, allowing for the retrieval of specific individuals using natural language descriptions. While effective for routine identification, existing methods struggle to address the complexities of real-world security, where detecting anomalies is paramount. Consequently, a new task, *text-based person anomaly search*, has been introduced. This task requires identifying pedestrians involved in both routine and anomalous activities. However, it faces a critical challenge: the "Pose-Semantic Gap." Distinct behaviors, such as *doing push-ups* versus *falling down*, often exhibit nearly identical skeletal geometries. Traditional pose-aware models, which rely heavily on geometric alignment, fail to discern these subtle semantic distinctions, leading to high false-positive rates.

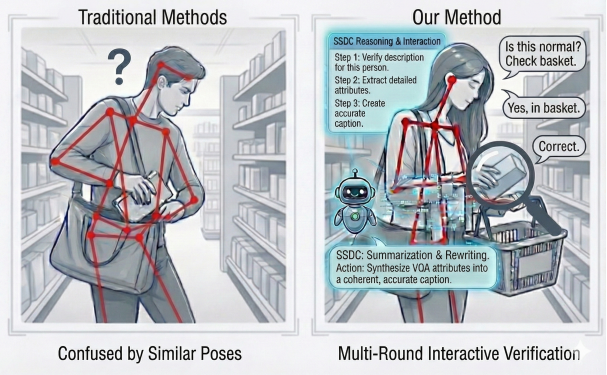


Figure 1: Illustration of the Pose-Semantic Gap. Traditional pose-aware methods (left) fail to distinguish semantically distinct actions with similar skeletal geometries. Our SSDC framework (right) bridges this gap through Multi-Round Interactive Verification using a collaborative agent workflow (Verify → Extract → Rewrite) to analyze fine-grained visual details and generate semantically accurate captions.

While Multimodal Large Language Models (MLLMs) possess the semantic reasoning capabilities necessary to resolve such ambiguities, deploying them directly on massive surveillance archives is computationally prohibitive. The sheer volume of video data demands high-throughput processing, yet MLLMs suffer from significant inference latency. This creates a dilemma: relying on lightweight models that sacrifice semantic precision for speed, or employing large models that offer accuracy but lack the efficiency required for real-time deployment.

To bridge this gap, we propose the Structure-Semantic Decoupled Cascade (SSDC) Framework.

*Corresponding author.



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 4040–4049

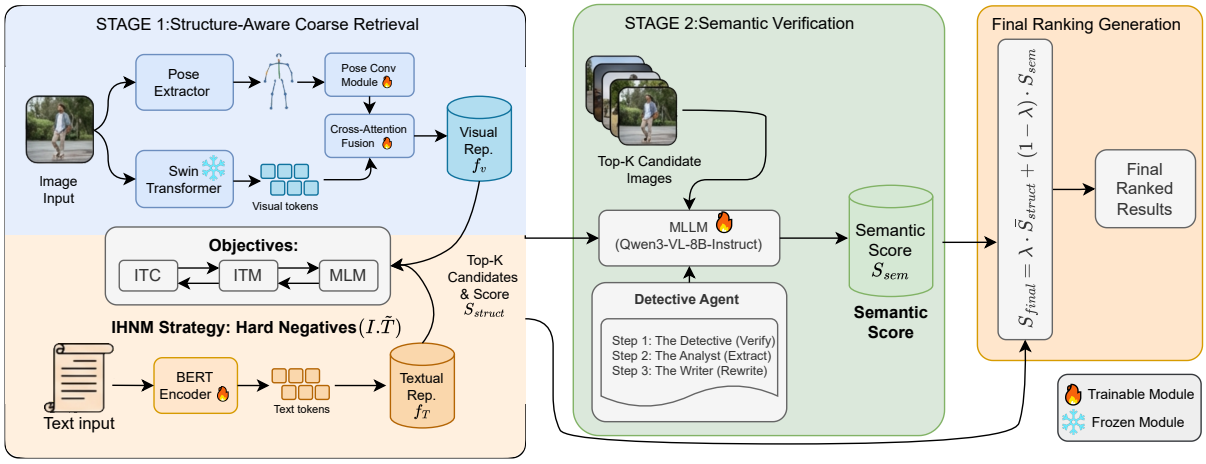


Figure 2: **Overall architecture of the SSDC framework .** The framework follows a coarse-to-fine pipeline : (1) Coarse Retrieval uses a lightweight model to filter the gallery based on structural similarity. (2) Semantic Verification introduces a specialized Detective Agent to scrutinize hard negatives. This agent employs Detective-style Prompting to resolve fine-grained ambiguities through multi-round reasoning and visual entailment. Finally, an Adaptive Fusion Mechanism integrates structural scores with the agent's semantic verdicts to produce the final ranking.

Our approach synergizes the efficiency of structural filtering with the reasoning power of large models. The retrieval process is decomposed into two synergistic stages: (1) A **Structure-Aware Coarse Retrieval** stage, where a lightweight Structure-Aware Coarse Retriever acts as a high-speed filter to recall a candidate pool based on structural similarity; and (2) A **Detective Squad Interaction** stage, where we replace monolithic verification with a novel multi-agent collaboration. Unlike previous methods, we orchestrate three specialized agents: a **Detective** for rapid binary filtering to discard obvious negatives, an **Analyst** for structured evidence extraction via a checklist, and a **Writer** for synthesizing fragmented evidence into a comprehensive, high-quality caption. Finally, this refined semantic description is used for precision re-ranking.

To support this task and validate our framework, we utilize the PAB benchmark (Yang et al., 2025b). This large-scale dataset comprises over 1 million image-text pairs, encompassing 1,600 anomaly types and 1,000 normal action types. In summary, our primary contributions are:

- We propose the SSDC Framework, a coarse-to-fine architecture that effectively bridges the Pose-Semantic Gap by decoupling structural filtering from semantic verification, balancing retrieval efficiency with reasoning depth.
- We design the **Detective Squad**, a multi-agent verification mechanism with three roles: Detective (Filter), Analyst (Extractor), and Writer (Integrator). Notably, this framework serves as a **plug-and-play** module that seamlessly adapts to various retrieval backbones.
- Extensive experiments on the PAB benchmark show that our method significantly outperforms SOTA baselines, validating our agent-based collaborative strategy.

## 2 Related Work

### 2.1 Text-Based Person Anomaly Search.

This task sits at the intersection of person retrieval and anomaly detection, requiring the localization of specific pedestrians based on fine-grained behavioral descriptions. In the domain of Text-Based Person Retrieval, methods have evolved from early global alignment (Zheng et al., 2020; Ding et al., 2021) to precise local matching via attention mechanisms (Wang et al., 2022; Shu et al., 2023). Recently, Vision-Language Pre-training (VLP) models like IRRA (Jiang and Ye, 2023b) and RaSa (Bai et al., 2023) have achieved state-of-the-art results by leveraging CLIP (Radford et al., 2021) for robust representation. Further, recent works have explored uncertainty modeling to handle noisy text-image correspondences (Xie et al., 2025a). *However*, these approaches primarily target static appearance attributes and overlook complex behavioral semantics, leading to a "Pose-Semantic Gap" where geometrically similar actions are misidentified. Conversely, in the realm of Person Anomaly Detection, traditional Video Anomaly Detection

(VAD) focuses on identifying statistical deviations in temporal sequences, often employing one-class classification (Zaheer et al., 2022; Flaborea et al., 2023) or weakly supervised ranking (Sultani et al., 2018). While effective for general monitoring, these methods lack the semantic flexibility to handle open-vocabulary natural language queries. Although recent works like UCA (Yuan et al., 2024) attempt to incorporate text, they operate at a coarse *video level* rather than the *instance level*. Consequently, neither domain alone solves the challenge of high-precision, instance-level anomaly retrieval.

### 2.2 Synergizing Structure and Semantics.

Integrating structural priors with semantic reasoning is a growing direction for disambiguating complex human behaviors. Pose-guided methods (Jing et al., 2020; Zhu et al., 2021) utilize keypoints to improve feature alignment. *However*, pose information alone remains ambiguous; distinct actions like *falling* and *push-ups* share identical skeletal geometries, leading to false positives when semantic context is absent. On the other hand, MLLMs demonstrate superior capabilities in semantic reasoning and visual entailment. While some studies utilize MLLMs for synthetic data generation (Yang et al., 2023; Xie et al., 2025b) or auxiliary supervision (Tan et al., 2024), directly deploying them for large-scale retrieval is computationally prohibitive due to high inference latency. Recent advancements have sought to alleviate these bottlenecks and enhance the reliability of large vision-language models in cross-modal retrieval through uncertainty-aware inference (Gong et al., 2026). Despite these efforts, existing pipelines often employ simple re-ranking strategies without explicitly modeling the structure-semantic discrepancy. In contrast, our Structure-Semantic Decoupled Cascade (SSDC) Framework uniquely synergizes these paradigms. We utilize a lightweight pose-aware model for efficient Coarse Retrieval and employ a specialized Detective Agent for Semantic Verification. This decoupled design effectively bridges the Pose-Semantic Gap, achieving a superior balance between retrieval accuracy and efficiency (Xie, 2026; Xie et al., 2026a,b).

## 3 Method

As illustrated in Figure 2, we propose the Structure-Semantic Decoupled Cascade (SSDC) Framework, a coarse-to-fine approach that bridges the "Pose-Semantic Gap" in text-based person anomaly search. Traditional single-stage models often struggle to distinguish semantically distinct behaviors with similar skeletal geometries. SSDC therefore decouples retrieval into two stages. First, a lightweight Structure-Aware Coarse Retriever filters the gallery by structural similarity to retrieve a candidate pool with high recall and low latency. Second, Detective Squad Interaction performs semantic verification. Unlike monolithic approaches, we use a collaborative multi-agent workflow: a Detective discards clear negatives, an Analyst extracts fine-grained visual evidence, and a Writer synthesizes it into a precise *new Query*. We then re-rank candidates using this refined description to improve semantic accuracy.

### 3.1 Structure-Aware Coarse Retriever

The primary goal of Stage I is to recall a high-quality candidate pool $\mathcal{C}$ from the gallery $\mathcal{G}$ ($|\mathcal{C}| \ll |\mathcal{G}|$). We adopt the Cross-Modal Pose-aware framework (Yang et al., 2025b) as the backbone.

**Pose-aware Representation Learning.** Anomalous behaviors often correspond to distinctive body configurations. We extract a pose map $P$ from the input image $I$ and encode it with a lightweight Pose Conv Module, while a Vision Transformer processes $I$ to produce patch-level embeddings $f_I$. We then fuse pose and image features via *Pose-aware Cross-Attention*: normalized pose features $f_P$ act as queries that attend to image features $f_I$, yielding a structure-enhanced representation $f_{CA}$:

$$f_{CA} = \text{Softmax}\left(\frac{(W_q f_P)(W_k f_I)^T}{\sqrt{d}}\right)(W_v f_I). \tag{1}$$

Finally, we obtain the visual embedding through a residual connection: $f_V = f_I + f_{CA}$.

**Training Objectives.** The coarse retriever is optimized using a hybrid objective: $\mathcal{L}_{stage1} = \mathcal{L}_{itc} + \mathcal{L}_{itm} + \mathcal{L}_{mlm}$. We employ Identity-based Hard Negative Mining to construct triplets where negative samples share the same identity but perform different actions, compelling the model to decouple action-specific semantics from identity appearance.

### 3.2 The Detective Squad Interaction

While Stage I efficiently filters structurally irrelevant samples, it lacks the reasoning depth to resolve fine-grained ambiguities. To bridge this gap, we introduce the **Detective Squad**, a multi-agent frame-

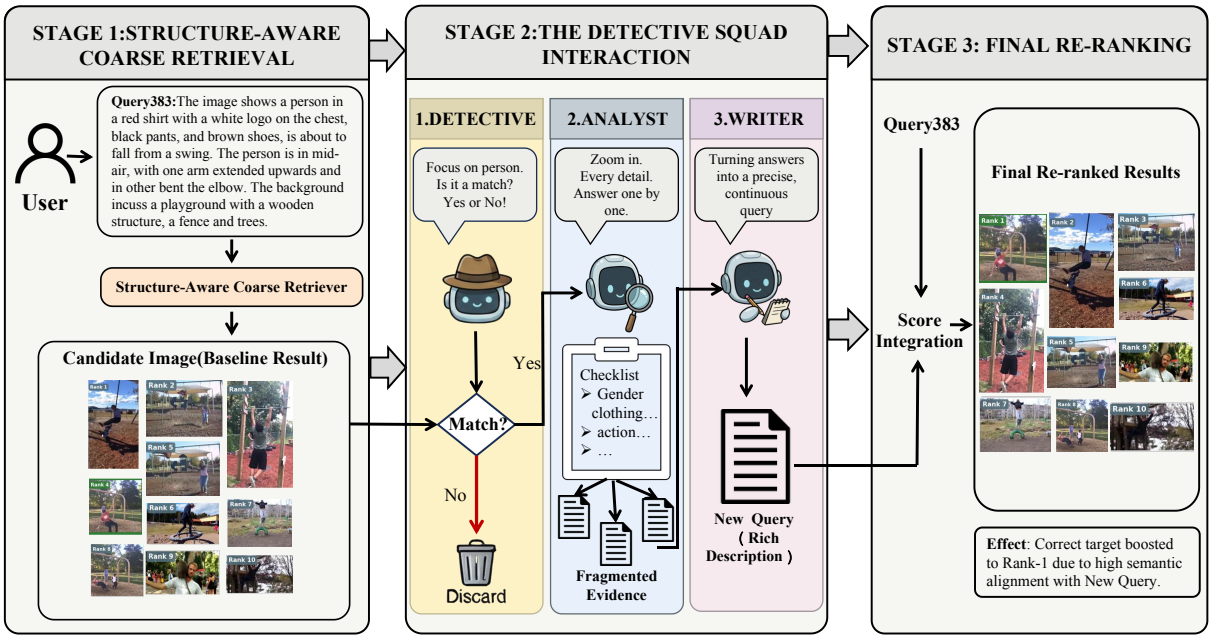


Figure 3: Overview of the proposed Detective Squad framework for person re-identification. The pipeline operates in a coarse-to-fine manner across three stages. Stage 1 utilizes a lightweight model for initial Structure-Aware Coarse Retrieval to obtain baseline candidates. Stage 2 introduces the novel multi-agent Detective Squad (Detective, Analyst, and Writer) to perform semantic verification, filter distractors, extract fine-grained details, and generate precise new captions. Stage 3 executes Final Re-ranking by fusing scores based on the generated captions, significantly boosting the rank of the target image.

work designed to conduct hierarchical interactive reasoning on the top-$K$ candidates.

### 3.2.1 Constructing the Detective Squad

General-purpose MLLMs often lack sensitivity to specific anomaly definitions. We bridge this domain gap through a Hard-Negative Aware Supervised Fine-Tuning (SFT) strategy.

**Mining Hard Negatives.** A competent detective must distinguish the true target from mimics. We simulate this by mining *structural hard negatives* from the frozen Stage I retriever. For a query $T$, we identify images $I^-$ that are ranked high due to structural similarity but are semantically incorrect. These tricky suspects force the agents to look beyond skeletal geometry. We subsequently aggregate these samples to construct a specialized dataset $\mathcal{D}_{sft}$, which serves as the foundation for training our agents to distinguish anomalies from routine actions.

**Role-Specific Instruction Tuning.** Instead of a simple binary task, we formulate verification as a *Collaborative Visual Reasoning* problem. We construct instruction templates that correspond to three specialized roles: (1) A **Detective** tailored for binary filtering (Match or Discard); (2) An **Analyst** trained to answer a fine-grained checklist (e.g., "Gender, Clothing, Action"); and (3) A **Writer** optimized to synthesize these details into coherent captions. We employ Low-Rank Adaptation (LoRA) to efficiently optimize the backbone, ensuring each agent adheres to its specific persona.

### 3.2.2 Detective-style Inference Chain

Standard captioning often yields hallucinations. To mitigate this, we propose a Detective-style Prompting mechanism that guides the squad's sequential reasoning process.

**Hypothesis Testing via Role-Playing.** We inject system prompts that assign specific personas to the agents, triggering a chain-of-thought verification. First, the **Detective** acts as a filter, asking "Is it a match? Yes or No!" to discard obvious distractors. Second, for surviving candidates, the **Analyst** performs a "physical examination," checking attributes against a 15-point checklist to extract fragmented evidence. Finally, the **Writer** acts as an integrator, synthesizing these verified details into a precise, continuous new caption $T_{new}$.

**Generative Semantic Scoring.** Unlike opaque probability logits, our method produces an interpretable semantic anchor. We verify the consis-

tency between the original query $T$ and the Writer's synthesized caption $T_{new}$ by computing their cosine similarity in the feature space:

$$S_{sem}(I_k, T) = \frac{E_{txt}(T) \cdot E_{txt}(T_{new})}{\|E_{txt}(T)\| \|E_{txt}(T_{new})\|} \quad (2)$$

where $E_{txt}$ denotes the frozen text encoder (e.g., from the CLIP or BERT backbone) used to extract semantic embeddings. This explicit text-to-text matching score $S_{sem}$ quantifies the semantic entailment, serving as the input for the subsequent fusion stage.

**Generative Semantic Re-ranking.** Instead of opaque probability scores, our method outputs explicit semantics. The *new Query* generated by the Writer serves as a semantic anchor. We compare it with the original query to compute a high-fidelity semantic score, bridging the pose-semantic gap through interpretable text generation rather than black-box logits.

#### 3.2.3 Efficiency-Aware Dynamic Inference

To apply the Detective Squad at scale, efficiency is critical. Inspired by cascade designs, we propose a Threshold-Gated Interaction mechanism.

**Dynamic Activation Strategy.** We observe that retrieval quality correlates with Stage I structural similarity. Candidates with low structural scores are likely irrelevant, making detailed analysis unnecessary. Therefore, the Detective agent serves as an efficient gatekeeper. It is activated only for top-ranked candidates. If the Detective outputs "No", the sample is immediately discarded. This strategy effectively filters out distinct negatives, allowing the computationally more expensive Analyst and Writer to focus strictly on ambiguous candidates .

### 3.3 Adaptive Fusion Mechanism

The final ranking is determined by fusing structural priors with semantic reasoning. The semantic score $S_{sem}$ is calculated by measuring the textual similarity between the user's original query $T$ and the *New Caption* $T_{new}$ synthesized by the Writer agent from the candidate image $I_k$:

$$S_{sem}(I_k, T) = \frac{E_{txt}(T) \cdot E_{txt}(T_{new})}{\|E_{txt}(T)\| \|E_{txt}(T_{new})\|} \quad (3)$$

where $E_{txt}$ denotes the frozen text encoder (e.g., BERT or CLIP text tower) used to extract semantic embeddings.

To balance efficiency and accuracy, we propose a threshold-gated fusion strategy. The final similarity score $S_{final}$ is computed as:

$$S_{final} = \begin{cases} \lambda \tilde{S}_{str} + (1-\lambda) S_{sem}, & \tilde{S}_{str} > \xi \\ \tilde{S}_{str}, & \text{else} \end{cases} \quad (4)$$

where $\tilde{S}_{str}$ is the min-max normalized structural similarity score from Stage I, and $\xi$ is the confidence threshold. This ensures that the computationally intensive semantic verification is reserved strictly for high-potential candidates.

## 4 Experiments

In this section, we conduct extensive experiments on public benchmarks to evaluate the effectiveness, superiority, and generalization of SSDC.

### 4.1 Experimental Setup

**Datasets.** We conduct experiments on the Pedestrian Anomaly Behavior (PAB) benchmark (Yang et al., 2024), a large-scale dataset comprising over 1 million image-text pairs covering 1,600 anomaly types and 1,000 normal action types. To evaluate robustness against environmental variations, we employ a Multi-weather Setting that simulates 10 distinct weather conditions, mimicking round-the-clock smart city scenarios. Finally, to assess out-of-distribution generalization, we evaluate on the UCC test set derived from the real-world UCF-Crime dataset (Sultani et al., 2018), which contains 5,320 unseen image-text pairs.

**Evaluation Metrics.** Following standard retrieval protocols (Li et al., 2017), we report Recall@K (R@1, R@5, R@10) and mAP. A successful retrieval requires the top-ranked result to perfectly align with the query in terms of appearance, action intent, and background context.

**Implementation Details.** All models are implemented in PyTorch.

In Stage I, we train the Structure-Aware Coarse Retriever on a single RTX 3090 GPU for 30 epochs with a batch size of 22 using AdamW; the learning rate decays linearly from $1 \times 10^{-4}$ to $1 \times 10^{-5}$. In Stage II (**Detective Squad Interaction**), we use a cloud instance with a single NVIDIA A6000 GPU and BF16 precision. We adopt Qwen3-VL-8B-Instruct (Bai et al., 2025a)as the unified backbone and fine-tune it with LoRA on our multi-role instruction dataset (9,000 samples) for 2 epochs. The dataset includes binary verification (Detective) and

| Method | Normal | | Wind | | Rain | | Snow | | Rain+Snow | | Dark | | Dark+Wind | | Dark+Rain | | Dark+Snow | | 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 |
| Baseline | 84.28 | 91.28 | 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 | 63.00 | 73.80 |
| IRRA (Jiang and Ye, 2023a) | 78.67 | 87.74 | 75.08 | 85.40 | 56.62 | 70.37 | 58.34 | 72.06 | 52.68 | 66.04 | 75.78 | 85.73 | 71.64 | 82.95 | 38.32 | 50.03 | 51.82 | 64.66 | 62.11 | 73.89 |
| CMP (Yang et al., 2025b) | 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 | 66.12 | 76.54 |
| RDE (Qin et al., 2024) | 76.74 | 86.12 | 72.24 | 83.13 | 49.90 | 64.40 | 53.03 | 66.69 | 46.21 | 59.71 | 71.79 | 82.74 | 67.54 | 79.24 | 31.55 | 41.54 | 46.21 | 58.47 | 54.81 | 66.99 |
| RDE + Detective Squad Interaction | 84.88 | 91.08 | 80.59 | 88.09 | 65.32 | 74.20 | 65.93 | 74.81 | 61.17 | 69.57 | 75.78 | 85.73 | 77.50 | 85.39 | 41.61 | 47.48 | 56.88 | 64.88 | 65.60 | 73.77 |
| IRRA + Detective Squad Interaction | 84.93 | 91.66 | 81.65 | 88.88 | 66.68 | 76.06 | 67.85 | 77.44 | **64.00** | **72.76** | 81.50 | 88.83 | 79.12 | 87.10 | **49.04** | **56.48** | 61.07 | 69.54 | 70.65 | 78.75 |
| **SSDC** | **87.01** | **92.74** | **84.78** | **91.22** | **67.90** | **76.68** | **69.36** | **78.68** | 62.29 | 71.33 | **82.86** | **90.10** | **80.59** | **87.88** | 45.10 | 51.47 | **61.78** | **70.03** | **71.30** | **78.90** |

Table 1: Robust text-based person anomaly retrieval results on PAB under multi-weather setting.

attribute checking (Analyst) instructions. We apply LoRA to all linear modules with rank 8, alpha 16, and dropout 0.05. We optimize with a cosine scheduler (peak learning rate $5.0 \times 10^{-5}$ , warmup ratio 0.1) and an effective batch size of 16. During inference, we use *Efficiency-Aware Dynamic Inference*: the Detective gates subsequent reasoning, and we fuse the final scores with $\lambda = 0.4$.

### 4.2 Comparison with State-of-the-Art.

Table 2 summarizes the quantitative performance of our SSDC framework alongside various SOTA methods, including general VLP models (e.g., CLIP (Radford et al., 2021), X-VLM (Zeng et al., 2022)) and specialized person retrieval methods (e.g., IRRA (Jiang and Ye, 2023b), RaSa (Bai et al., 2023)).

Under the zero-shot setting, general models exhibit limited adaptability; for instance, CLIP only reaches 47.57% R@1. This confirms that standard semantic alignment struggles to bridge the specialized "Pose-Semantic Gap" in anomaly search. When fine-tuned on the 0.1M PAB dataset, our Stage I pose-aware retriever (CMP (Yang et al., 2025b)) already achieves 83.06% R@1 and 90.41% mAP. It outperforms established baselines such as X-VLM (81.95% R@1) and RaSa (80.79% R@1), validating the effectiveness of incorporating structural pose priors.

With the full 1M training set, our complete SSDC framework sets a new state-of-the-art across all primary metrics, reaching **87.01%** R@1 and **92.74%** mAP. This significantly surpasses both the strong IRRA baseline (78.67% R@1) and the standalone CMP model (84.93% R@1). Notably, integrating the **Detective Squad Interaction** into a simpler backbone (IRRA + Detective Squad) enhances the R@1 to 83.92%. This demonstrates that our collaborative verification stage effectively resolves hard ambiguity cases through fine-grained evidence extraction, leading to the best overall performance regardless of the base retriever.

**Robustness to Environmental Variations.** Real-world surveillance often suffers from visual degradation. We evaluate model robustness across 10 distinct weather conditions (e.g., rain, snow, dark), as shown in Table 1. SSDC demonstrates consistent performance gains across all scenarios, proving that the Detective Squad's hierarchical reasoning provides resilience even when visual structural cues are compromised.

**Out-of-Distribution Generalization.** To evaluate generalization, we test on the unseen UCC dataset. Table 3 shows that our model achieves **59.45%** R@1 and **45.25%** mAP, significantly outperforming baselines trained on the same data (e.g., RaSa at 54.12% R@1). This demonstrates that our multi-agent workflow allows the model to learn generalized representations of anomaly intents rather than overfitting to specific dataset biases.

### 4.3 Ablation Studies

We conduct a comprehensive ablation study to validate the effectiveness of each module within the Detective Squad and the necessity of our fine-tuning strategy (see Table 4). The baseline Structure-Aware Coarse Retriever achieves an R@1 of 84.28%. Regarding the contribution of individual agents, introducing the **Analyst** yields a substantial performance gain (↑ 1.92% R@1), significantly outperforming the **Writer**-only configuration (↑ 0.40%). This empirical evidence underscores that the core bottleneck in bridging the Pose-Semantic Gap is the lack of *explicit visual evidence extraction*. While the Writer can smooth the textual query, it struggles to generate new semantic information without the Analyst's fine-grained observations. Furthermore, the full collaborative squad achieves the peak performance of **87.21%**, demonstrating a strong synergy where the combined agents outperform their individual contributions. Finally, comparing the full squad with LoRA against the zero-shot prompting baseline, we observe a clear performance drop of 1.42% (87.21% → 85.79%) when LoRA is removed. This indicates that while modern MLLMs possess general knowledge, LoRA fine-tuning is indispensable for aligning the model's attention with specific anomaly def-

| Method | Ref. | Image Enc. | Text Enc. | R@1 | R@5 | R@10 | mAP |
|---|---|---|---|---|---|---|---|
| *Zero-shot Performance* | | | | | | | |
| MRA (Yang et al., 2025a) | CVPR'25 | Swin-B | BERT-Base | 9.91 | 23.66 | 31.45 | 17.15 |
| RaSa (Bai et al., 2023) | IJCAI'23 | ViT-B/16 | BERT-Base | 21.74 | 27.30 | 27.96 | 24.35 |
| WoRA (Sun et al., 2025) | WWW '25 | Swin-B | BERT-Base | 22.25 | 45.91 | 53.54 | 33.39 |
| APTM (Yang et al., 2023) | MM'23 | Swin-B | BERT-Base | 22.90 | 45.80 | 52.38 | 33.56 |
| CAMeL (Yu et al., 2025) | CVPR'25 | SG-Former | BERT-Base | 24.47 | 50.00 | 58.75 | 36.75 |
| IRRA (Jiang and Ye, 2023b) | CVPR'23 | ViT-B/16 | Transformer | 30.59 | 59.61 | 68.91 | 44.41 |
| CLIP (Radford et al., 2021) | ICML'21 | ViT-B/16 | Transformer | 47.57 | 81.55 | 89.03 | 62.73 |
| X-VLM (Zeng et al., 2022) | ICML'22 | Swin-B | BERT-Base | 71.94 | 97.78 | 98.99 | 83.96 |
| RDE (Qin et al., 2024) | CVPR'24 | ViT-B/16 | Transformer | 41.30 | 69.01 | 77.35 | 54.36 |
| IRRA + Detective Squad Interaction | - | ViT-B/16 | Transformer | 60.57 | 75.08 | 81.45 | 68.00 |
| RDE + Detective Squad Interaction | - | ViT-B/16 | Transformer | 63.55 | 78.56 | 81.50 | 71.00 |
| *Fine-tuned with 0.1M PAB* | | | | | | | |
| MRA (Yang et al., 2025a) | CVPR'25 | Swin-B | BERT-Base | 70.53 | 94.69 | 97.47 | 81.59 |
| APTM (Yang et al., 2023) | MM'23 | Swin-B | BERT-Base | 72.14 | 95.30 | 97.17 | 82.78 |
| CAMeL (Yu et al., 2025) | CVPR'25 | SG-Former | BERT-Base | 74.30 | 96.79 | 98.84 | 84.20 |
| WoRA (Sun et al., 2025) | WWW '25 | Swin-B | BERT-Base | 74.47 | 96.82 | 98.48 | 84.60 |
| IRRA (Jiang and Ye, 2023b) | CVPR'23 | ViT-B/16 | Transformer | 76.39 | 97.62 | 99.14 | 86.33 |
| CLIP (Radford et al., 2021) | ICML'21 | ViT-B/16 | Transformer | 77.60 | 98.84 | **99.75** | 87.35 |
| RaSa (Bai et al., 2023) | IJCAI'23 | ViT-B/16 | BERT-Base | 80.79 | 98.89 | 99.65 | 89.20 |
| X-VLM (Zeng et al., 2022) | ICML'22 | Swin-B | BERT-Base | 81.95 | 98.84 | 99.19 | 89.86 |
| CMP (Yang et al., 2025b) | ICCV'25 | ViT-B/16 | BERT-Base | 83.06 | 98.89 | 99.49 | 90.41 |
| *Fine-tuned with 1M PAB* | | | | | | | |
| IRRA (Jiang and Ye, 2023b) | CVPR'23 | ViT-B/16 | Transformer | 78.67 | 97.98 | 98.94 | 87.74 |
| CMP (Yang et al., 2025b) | ICCV'25 | ViT-B/16 | BERT-Base | 84.93 | **99.09** | **99.75** | 91.66 |
| RDE (Qin et al., 2024) | CVPR'24 | ViT-B/16 | Transformer | 76.74 | 96.97 | 98.38 | 86.12 |
| RDE + Detective Squad Interaction | - | ViT-B/16 | Transformer | 84.88 | 98.13 | 98.69 | 91.08 |
| IRRA + Detective Squad Interaction | - | ViT-B/16 | Transformer | 83.92 | 98.33 | 99.09 | 90.60 |
| **SSDC** | - | **ViT-B/16** | **BERT-Base** | **87.21** | **99.09** | **99.75** | **92.87** |

Table 2: Quantitative results of our proposed method and compared methods on the PAB benchmark. **Bold** indicates the best result, and underlined indicates the second best.

| Method | R@1 | R@5 | R@10 | mAP |
|---|---|---|---|---|
| APTM (Yang et al., 2023) | 27.86 | 40.41 | 46.77 | 22.61 |
| CLIP (Radford et al., 2021) | 51.60 | 68.31 | 76.43 | 43.05 |
| X-VLM (Zeng et al., 2022) | 52.33 | 66.73 | 72.54 | 40.87 |
| RaSa (Bai et al., 2023) | 54.12 | 70.32 | 75.96 | 39.71 |
| IRRA (Jiang and Ye, 2023b) | 40.28 | 57.24 | 65.98 | 33.53 |
| RDE (Qin et al., 2024) | 32.69 | 48.55 | 56.64 | 27.42 |
| CMP (Yang et al., 2025b) | 55.23 | 71.67 | 77.99 | 44.35 |
| RDE + Detective Squad Interaction | 41.58 | 48.52 | 56.39 | 28.20 |
| IRRA + Detective Squad Interaction | 49.00 | 58.84 | 67.08 | 34.69 |
| SSDC | **59.45** | **72.68** | **78.30** | **45.25** |

Table 3: Comparisons with existing methods in OOD setting. The unseen test set UCC is extracted from the UCF-Crime (Sultani et al., 2018) dataset.

| No. | Detective | Analyst | Writer | LoRA | R@1 | mAP |
|---|---|---|---|---|---|---|
| 1 | ✗ | ✗ | ✗ | ✗ | 84.28 | 91.28 |
| 2 | ✓ | ✗ | ✗ | ✓ | 84.28 | 91.28 |
| 3 | ✓ | ✓ | ✗ | ✓ | 86.20 | 92.34 |
| 4 | ✓ | ✗ | ✓ | ✓ | 84.68 | 91.54 |
| 5 | ✓ | ✓ | ✓ | ✗ | 85.79 | 92.12 |
| 6 | ✓ | ✓ | ✓ | ✓ | **87.21** | **92.87** |

Table 4: **Ablation study of the Detective Squad Interaction.** We analyze the contribution of each agent role and the impact of the LoRA fine-tuning strategy.

initions and ensuring strict adherence to the multi-agent workflow.

### 4.4 Impact of Foundation Model Selection.

We justify utilizing Qwen3-VL-8B(Bai et al., 2025a) as the unified backbone by benchmarking it against the Qwen2.5 series and the state-of-the-art InternVL3.5-8B(Wang et al., 2025). As shown in Table 5, while Qwen2.5-VL (Bai et al., 2025b)provides a decent baseline, it lacks the deep reasoning required for subtle anomalies. InternVL3.5-8B proves to be a strong competitor with impressive visual understanding (84.65% R@1). However, Qwen3-VL-8B achieves the best overall performance (84.73% R@1). We attribute this superiority to its balanced proficiency in both *visual chain-of-thought* (crucial for the Analyst) and *complex instruction following* (crucial for the Writer). Consequently, we select Qwen3-VL-8B as the optimal single-model engine to drive our collaborative squad.

| Stage-wise Model Selection | | | Performance | |
|---|---|---|---|---|
| Detective | Analyst | Writer | R@1 | mAP |
| Qwen2.5-VL-7B | Qwen2.5-VL-7B | Qwen2.5-7B | 85.69 | 92.03 |
| InternVL3.5-8B | InternVL3.5-8B | InternVL3.5-8B | 86.10 | 92.19 |
| Qwen3-VL-8B | Qwen3-VL-8B | Qwen3-VL-8B | **86.20** | **92.34** |

Table 5: Ablation study on Foundation Model Selection.

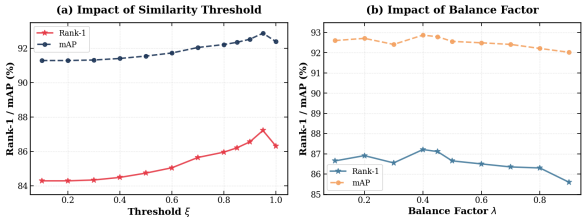


Figure 4: Parameter sensitivity analysis of SSDC.

### 4.5 Efficiency Analysis

We analyze the trade-off between accuracy and inference cost. Directly applying the multi-agent Detective Squad to the full 1M+ gallery is computationally prohibitive, so we use a cascade strategy for real-time deployment. First, the lightweight CMP model filters out over 99.9% of candidates and invokes the squad only for top-ranked samples that satisfy a high-confidence threshold ($S_{str} > 0.95$). Second, within the squad, we use *Efficiency-Aware Dynamic Inference*: the Detective acts as a gatekeeper, and the computationally intensive Analyst and Writer run *if and only if* the Detective verifies a "Match". This design focuses expensive reasoning on the most ambiguous samples, reducing latency by orders of magnitude while preserving the accuracy of exhaustive re-ranking.

### 4.6 Parameter Sensitivity Analysis

**Impact of Similarity Threshold $\xi$.** As shown in Figure 4(a), we analyze the threshold $\xi$ which governs the activation of the Detective Squad. We observe a steady improvement in Rank-1 accuracy as the threshold increases, peaking at $\xi = \mathbf{0.95}$. This trend validates our Efficiency-Aware strategy: by setting a high threshold, we effectively filter out structurally irrelevant noise, ensuring that the computationally intensive semantic verification is reserved strictly for ambiguous, high-value candidates that genuinely require fine-grained scrutiny.

**Impact of Balance Factor $\lambda$.** We further analyze the fusion weight $\lambda$, which balances the structural priors ($S_{str}$) and semantic scores ($S_{sem}$). As illustrated in Figure 4(b), performance maximizes at $\lambda = \mathbf{0.4}$, implying a higher reliance on the semantic score ($1 - \lambda = 0.6$). This confirms that while structural alignment is essential for initial recall, the detailed reasoning provided by the Detective Squad serves as the decisive factor in resolving the Pose-Semantic Gap. The results remain robust within the range $[0.3, 0.5]$, leading us to set $\lambda = 0.4$ as the default.

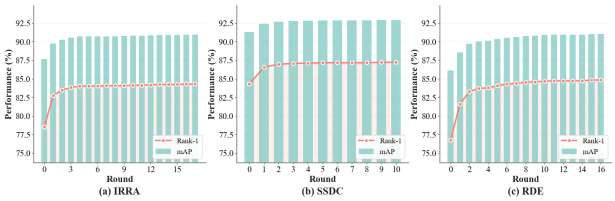


Figure 5: Evolution of Rank-1 and mAP performance versus interaction rounds for IRRA, SSDC, and RDE. Round 0 denotes the baseline result without the Detective Squad. Subsequent rounds represent the iterative refinement cycles.

### 4.7 Iterative Reasoning Analysis

This section further explores the efficacy of the **Detective Squad** in resolving semantic ambiguities through iterative refinement. We report the performance changes in terms of mAP and Rank-1 (R@1) in Figure 5, where **one interaction round** is defined as a complete inference cycle passing through all three agents . Experimental results indicate that the most significant improvements in both mAP and R@1 occur in the initial rounds ($\leq 2$). This is attributed to the **Detective Squad** completing its primary verification loop: the **Detective** first filters distinct hard negatives, the **Analyst** extracts fine-grained evidence, and the **Writer** synthesizes a precise caption to bridge the Pose-Semantic Gap. As the number of refinement cycles increases, the performance gains gradually stabilize. This confirms that the initial collaborative pass of the squad effectively resolves the majority of ambiguities, while subsequent rounds provide marginal semantic polishing.

## 5 Conclusion

In this paper, we address the challenge of Text-based Person Anomaly Search by leveraging the large-scale Pedestrian Anomaly Behavior (PAB) benchmark (Yang et al., 2025b) to bridge the gap between synthetic training and real-world evaluation. To resolve the inherent "Pose-Semantic Gap," we propose the Structure-Semantic Decoupled Cascade (SSDC) framework. By synergizing a

lightweight pose-aware retriever with a multi-agent Detective Squad, our approach effectively resolves fine-grained ambiguities that traditional methods miss. Extensive experiments confirm that SSDC establishes new state-of-the-art performance, demonstrating remarkable robustness across diverse environmental conditions.

## 6 Limitations

While our SSDC framework achieves state-of-the-art performance, we identify two primary limitations. First, as a cascade architecture, the final retrieval performance is upper-bounded by the Stage I coarse retriever. If the relevant target is not recalled within the initial candidate pool, the Detective Squad cannot recover it, leading to inevitable misses. Second, despite the efficiency-aware gating mechanism, the deployment of an 8B-parameter MLLM still requires substantial GPU resources . This computational overhead currently restricts the framework's applicability on resource-constrained edge devices, such as standalone surveillance cameras, where lightweight efficiency is paramount.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. U24A20326, and the State Grid Zhejiang Electric Power Cooperation Technology Project (Grant Number: B311DS240012).

## References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

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

Zefeng Ding, Changxing Ding, Zhiyin Shao, and Dacheng Tao. 2021. Semantically self-aligned network for text-to-image part-aware person re-identification. *arXiv preprint arXiv:2107.12666*.

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

Tianxiang Gong, Shiqi Gao, Qi Song, Qingyun Sun, Haoyi Zhou, and Jianxin Li. 2026. Towards reliable multimodal intelligence via uncertainty-aware inference. *Chinese Journal of Electronics*, pages 1–16. Early Access.

D. Jiang and M. Ye. 2023a. 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*, pages 2787–2797.

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

Ya Jing, Chenyang Si, Junbo Wang, Wei Wang, Liang Wang, and Tieniu Tan. 2020. Pose-guided multi-granularity attention network for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11189–11196.

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

Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. 2024. Noisy-correspondence learning for text-to-image person re-identification. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

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

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

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

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

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

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

Zequn Xie. 2026. Conquer: Context-aware representation with query enhancement for text-based person search. *arXiv preprint arXiv:2601.18625*.

Zequn Xie, Haoming Ji, Chengxuan Li, and Lingwei Meng. 2025a. Dynamic uncertainty learning with noisy correspondence for text-based person search. *arXiv preprint arXiv:2505.06566*.

Zequn Xie, Xin Liu, Boyun Zhang, Yuxiao Lin, Sihang Cai, and Tao Jin. 2026a. Hvd: Human vision-driven video representation learning for text-video retrieval. *arXiv preprint arXiv:2601.16155*.

Zequn Xie, Chuxin Wang, Yeqiang Wang, Sihang Cai, Shulei Wang, and Tao Jin. 2025b. Chat-driven text generation and interaction for person retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5259–5270.

Zequn Xie, Boyun Zhang, Yuxiao Lin, and Tao Jin. 2026b. Delving deeper: Hierarchical visual perception for robust video-text retrieval. *arXiv preprint arXiv:2601.12768*.

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

Shuyu Yang, Yaxiong Wang, Li Zhu, and Zhedong Zheng. 2024. Beyond walking: A large-scale image-text benchmark for text-based person anomaly search. *arXiv preprint arXiv:2411.17776*.

Shuyu Yang, Yaxiong Wang, Li Zhu, and Zhedong Zheng. 2025b. Beyond walking: A large-scale image-text benchmark for text-based person anomaly search. In *ICCV*.

Shuyu Yang, Yinan Zhou, Yaxiong Wang, Yujiao Wu, Li Zhu, and Zhedong Zheng. 2023. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 2023 ACM on Multimedia Conference*.

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

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

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

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

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

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

Aichun Zhu, Zijie Wang, Yifeng Li, Xili Wan, Jing Jin, Tian Wang, Fangqiang Hu, and Gang Hua. 2021. Dssl: Deep surroundings-person separation learning for text-based person retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 209–217.