Causal Intervention with LMMs: Generative Text Facilitates Fine-Grained Pedestrian Anomaly Behavior Retrieval

Anonymous ACL submission

Abstract

Pedestrian anomaly behavior retrieval requires disambiguating subtle anomalous actions from routine ones using natural language. Existing cross-modal alignment methods often struggle with confounding biases: visual entanglement (dominant environments overshadowing target actions) and textual biases (limited stylistic diversity). These spurious correlations lead to erroneous matches among visually similar but semantically distinct samples. To address this, we reformulate the task from a causal intervention perspective. We propose a dual generative framework leveraging Large Multimodal Models (LMMs) to break environment-action and style-text confounders. Specifically, we synthesize causally meaningful hard negatives to decouple actions from environments, and augment positive queries with stylistically diverse rephrasings to eliminate narrative dependency. Extensive experiments on the Pedestrian Anomaly Behavior (PAB) benchmark yield consistent improvements of +1.71% in Recall@1 and +1.01% in mAP. Furthermore, zero-shot evaluation on the out-of-distribution UCC dataset demonstrates markedly enhanced generalization, achieving +3.23% and +2.46% gains in Recall@1 and mAP, respectively.

1 Introduction

Pedestrian anomaly behavior retrieval (Li et al., 2017; Zheng and Zheng, 2024) aims to retrieve specific pedestrians exhibiting anomalous activities (*e.g.*, " falling " or " fighting ") from large-scale video galleries based on natural language descriptions. Unlike generic object or person retrieval (Philbin et al., 2007, 2008), which primarily matches static attributes such as clothing or identity, this task requires a fine-grained (Chen et al., 2023a) understanding of both visual appearance and dynamic behavioral semantics, disentangling subtle anomalous actions from routine ones within complex scenes. This capability holds significant real-world importance in intelligent public safety systems. By enabling rapid, text-guided localization of anomalous events (*e.g.*, detecting fights, falls, collisions, or suspicious behaviors in crowded areas, transportation hubs, or urban environments), such retrieval supports timely intervention to prevent crimes, mitigate accidents, enhance emergency response, and improve overall security monitoring (Yang et al., 2025b). The recent introduction of the Pedestrian Anomaly Behavior (PAB) benchmark (Yang et al., 2025b) addresses the critical need for dedicated large-scale resources in this domain, providing extensive image-text pairs that unify routine and anomalous pedestrian actions to facilitate robust training and evaluation of cross-modal retrieval models.

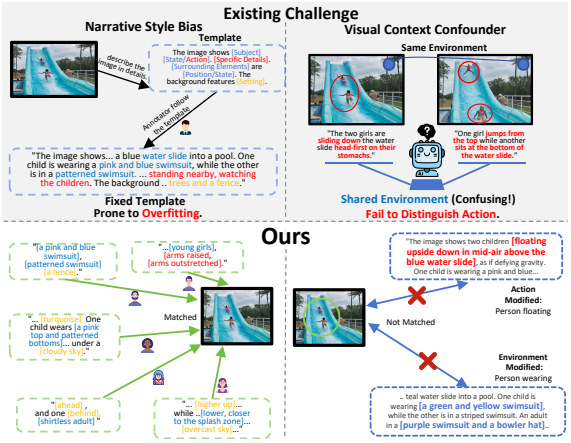

Figure 1: **Motivation.** ((*Top*) Existing methods suffer from confounding biases: limited narrative diversity causes textual overfitting, while environmental entanglement misleads models into global scene matching rather than fine-grained action distinction. (*Bottom*) Our LMM-based causal intervention tackles this by synthesizing diverse positives to break narrative dependencies, and counterfactual hard negatives to decouple actions from environments, prioritizing true behavioral cues over spurious artifacts.

While recent VLMs (Zhang et al., 2024a) excel

on general benchmarks, their reliance on statistical correlations proves unreliable for high-stakes anomaly retrieval (Chowdhury et al., 2024). We identify two critical limitations hindering existing models. First, standard PAB (Yang et al., 2025b) annotations suffer from semantic sparsity, introducing a narrative style bias. As shown in Figure 1, template-induced rigidity causes models to overfit specific sentence patterns rather than capturing wording-agnostic, invariant action semantics. Second, severe causal entanglement acts as a visual context confounder. Since actions frequently co-occur with specific environments (*e.g.*, "`sliding down`" and "`water slide`" in Figure 1), models learn spurious correlations. They shortcut to associate environments directly with anomaly labels, neglecting actual causal behavioral cues. Consequently, the combination of narrative and visual biases leads to poor generalization, particularly in Out-Of-Distribution (OOD) scenarios.

To tackle these confounding effects, we shift from observational likelihood matching to causal intervention. Rather than modifying complex architectures, we propose a data-centric dual generative causal intervention framework to deconfound the learning objective (Louizos et al., 2017) via two strategic mechanisms. First, to eliminate narrative style bias, we implement Semantic Intervention by constructing causally consistent positives via Multi-Caption Augmentation. This prevents visual-text alignment from overfitting fixed templates, enforcing robust and invariant representations. Second, to eliminate the visual context confounder, we perform Counterfactual Intervention. Leveraging Large Multimodal Models (LMMs), we synthesize two types of hard negatives: (i) ***Action-Specific Counterfactuals***, which alter motion verbs (*e.g.*, changing `sliding down` " to `jumping` ") within identical contexts to isolate the *Action*; and (ii) ***Environment-Specific Counterfactuals***, which swap visual attributes (*e.g.*, clothing, background) while preserving actions to break the spurious *Environment-Label* link. Contrasting queries against these causally consistent positives and counterfactual negatives compels the model to encourages disentanglement between behavioral cues and environmental noise. Finally, evaluations on the PAB (Yang et al., 2025b) benchmark and an unseen OOD test set, UCC (Yang et al., 2025b) (extracted from UCF-Crime (Sultani et al., 2018)), demonstrate that our strategy significantly improves retrieval accuracy and generalization over observational baselines. To sum up, our main contributions are as follows:

- We analyze fine-grained pedestrian anomaly retrieval from a causal perspective, identifying two overlooked confounding biases: visual environment-action entanglement and textual narrative style bias. These biases commonly cause mismatches among visually similar yet behaviorally distinct samples.

- We propose an architecture-agnostic generative causal intervention framework leveraging LMMs to synthesize causally decoupled hard negatives (action-/environment-specific) and stylistically diverse positives. This forces the model to learn invariant behavioral cues over spurious correlations.

- Extensive experiments on the PAB benchmark (Yang et al., 2025b) and the OOD UCC test set validate our method's superiority. It achieves 89.73% R@1 and 94.56% mAP on PAB (with +1.71% and +1.01% relative gains), and yields even more pronounced zero-shot improvements on UCC (+3.23% R@1, +2.46% mAP), exhibiting exceptional generalization and transferability.

2 Related Work

Pedestrian Anomaly Behavior Retrieval. This task (Chen et al., 2020; Gan et al., 2020; Li et al., 2020b,a, 2019) aims to locate specific pedestrian images using natural language queries. Unlike general cross-modal retrieval (Duan et al., 2023; Wang et al., 2020; Yao et al., 2023) targeting broad semantics, it requires capturing subtle visual cues like clothing, accessories, and crucial motion details.

Existing approaches (Bai et al., 2023; Han et al., 2021; Jiang and Ye, 2023; Yang et al., 2025b) generally utilize dual-stream (Mao et al., 2021; Wu et al., 2024) or single-stream (Khan et al., 2022) architectures. While dual-stream methods enable efficient large-scale retrieval (Chen et al., 2023b) via independent embedding spaces, they struggle with fine-grained region-word alignments. Conversely, single-stream methods mitigate this by modeling dense token-level interactions, significantly improving accuracy at the cost of higher computational complexity.

Generalizable VLP models (Li et al., 2021; Radford et al., 2021; Zhang et al., 2024b; Guo et al., 2023; Luo et al., 2023) are increasingly adapted

to pedestrian retrieval (Yang et al., 2025b), alongside LLMs (Li et al., 2022, 2023a) and LMMs (Liu et al., 2023; Achiam et al., 2023; Dai et al., 2023) synthesizing diverse training data to enrich semantic spaces. However, existing methods (Niu et al., 2024; Wang et al., 2024; Ramshetty et al., 2023) focus on static attributes and neglect behavior-context couplings, struggling to distinguish subtle actions from backgrounds. In contrast, our framework employs fine-grained counterfactual mining to explicitly deconfound anomaly behaviors from environmental and stylistic biases.

Causal Inference in Vision-Language Learning. Causal inference (Pearl, 2010, 2009) eliminates spurious correlations (*e.g.*, associating pools with falling) that cause out-of-distribution (OOD) failures. To overcome confounding biases, Structural Causal Models (SCMs) and the backdoor criterion (Neuberg, 2003) block non-causal paths via interventions $P(Y|do(X))$, forcing models to disentangle causal features via counterfactual reasoning. Recently, causal interventions have improved multimodal tasks (Su et al., 2025; Li et al., 2023b), including VQA (Niu et al., 2021; Salemi et al., 2023), Scene Graph Generation, and image-text matching (Liu et al., 2021). Extending this to pedestrian anomaly retrieval, we identify narrative style and background context as key confounders. To establish robust causal links, we synthesize counterfactuals via generative hard negative mining, altering specific actions or environments. Unlike prior works, this approach simultaneously mitigates stylistic and environmental biases.

3 Methodology

3.1 Problem Formulation & Causal Analysis

Standard Formulation. Let $\mathcal{D} = \{(I_i, T_i)\}_{i=1}^{N}$ denote a fine-grained anomaly retrieval dataset. The goal is to learn a mapping function $f(I, T)$ that aligns the visual manifestation I with the textual description T for semantic similarity ranking. However, in fine-grained scenarios, representation learning often suffers from poor generalization due to visual and textual confounding factors.

Structural Causal Model. To diagnose this failure, we formulate the data generation mechanism as a Directed Acyclic Graph (DAG) (Figure 2), with variables defined as follows:

- E (**Visual Context Confounder**): Environmental setting dictating I's background ($E \to I$).

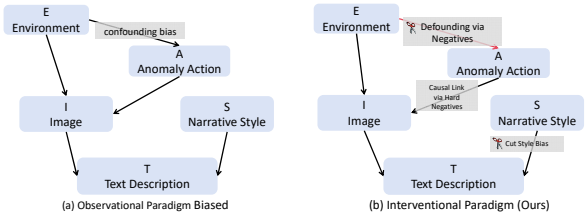

Figure 2: The Structural Causal Model for Fine-Grained Anomaly Retrieval. (a) **Observational Paradigm**: Visual context E introduces spurious correlation ($E \to A$), while narrative style S biases text T ($S \to T$). (b) **Interventional Paradigm (Ours)**: We cut the style bias ($S \nrightarrow T$) via Multi-Caption Augmentation and break the context confounding ($E \nrightarrow A$) via Counterfactual Intervention, thereby enforcing the true causal link ($A \to I$).

- A (**Anomaly Action**): Core semantic content dictating I's foreground ($A \to I$).
- I (**Image**): Governed by both action and environment ($A \to I \leftarrow E$).
- S (**Narrative Style Bias**): Annotators' linguistic patterns (*e.g.*, fixed templates).
- T (**Text**): Governed by both image and narrative style ($S \to T \leftarrow I$).

The Observational Paradigm (Biased). As illustrated in Figure 2(a), the observational data exhibits a strong confounding bias ($E \to A$). Certain actions are statistically bound to specific environments (*e.g.*, "sliding down" predominantly occurs on a "water slide"). This spurious association opens a backdoor path $A \leftarrow E \to I$. Consequently, the model exploits environmental cues (E) as shortcuts for action recognition, confounding the true behavioral signal ($A \to I$). Simultaneously, the textual query T suffers from narrative style bias ($S \to T$). Due to annotation sparsity, T is rigidly governed by I and S. Lacking diverse expressions, the model overfits to specific linguistic templates, failing to isolate core action semantics from surface-level phrasing.

The Interventional Paradigm (Ours). To achieve robust retrieval, we propose a Generative Causal Intervention framework (Figure 2(b)) to eliminate these biases. First, we implement the intervention $do(A)$ by synthesizing action-specific counterfactuals (associating context E with a conflicting action A'). This effectively severs the confounding link $E \to A$ and deprives the model of the environmental prior $P(A|E)$. Thus, the model is compelled to shift its attention from global scene

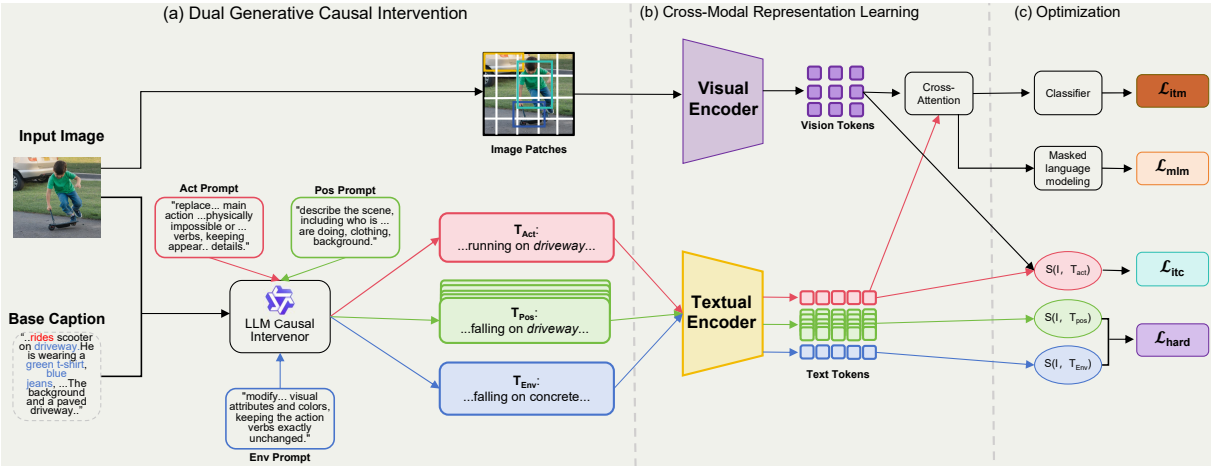

Figure 3: Overview of our Generative Causal Intervention framework. **(a) Dual Generative Causal Intervention**: An LMM synthesizes multi-style positives and counterfactual hard negatives to sever narrative style and context biases. **(b) Cross-Modal Representation Learning**: Visual and textual inputs are processed via a dual-stream encoder and cross-attention fusion. **(c) Optimization**: Alongside standard objectives, our proposed $\mathcal{L}_{hard}$ optimizes a disentangled metric space by pulling causally consistent positives toward the image anchor while explicitly pushing away action and environment confounders.

statistics to localized motion cues, establishing a robust causal link $A \rightarrow I$. Second, we mitigate style bias via Multi-Caption Augmentation. By using LMMs to diversify linguistic patterns in T, we approximate the marginalization over the latent distribution of S. This breaks the deterministic dependency $S \rightarrow T$, ensuring the visual-text alignment is driven by invariant action semantics A rather than template-induced rigidity.

3.2 Semantic Intervention via Multi-Caption Augmentation

Motivation. As analyzed in the Introduction, the standard annotations suffer from Narrative Style Bias ($S \rightarrow T$). The template-induced rigidity limits the semantic embedding space, causing the model to overfit to specific sentence patterns rather than capturing the invariant action semantics. To cut the link $S \rightarrow T$ and deconfound the narrative style, we implement Semantic Intervention.

Multi-Caption Augmentation. We leverage LMMs to generate a diverse set of captions $\mathcal{T}_{pos} = \{T^1, \ldots, T^K\}$ for each image. To break the fixed narrative structure of the original dataset, we design a diverse prompting strategy that varies the instructional focus while keeping the semantic content invariant. Specifically, we utilize the core prompt: "*Based strictly on the visual content, generate a concise and accurate caption following this guidance: '[template]'. Describe only what is clearly visible. Do not invent details or repeat information.*" To effectively marginalize the style variable S, we cycle through four distinct guidance templates to perturb the narrative perspective:

- *Scene-focused*: "Describe the scene, including who is present, what they are doing, their clothing, and the background."
- *Event-focused*: "What is happening? Mention the location, the people, their actions, attire, and key background details."
- *Summary-focused*: "Summarize the image: environment, number of people, their activity, clothing, and surroundings."
- *Setting-focused*: "Give a clear description of the setting, the individuals, their actions, and visible background elements."

By training on this expanded set, we effectively marginalize over the variable S:

$$\mathbb{E}_S[P(T|I,S)] \approx P(T|I). \quad (1)$$

This operation computationally severs the dependency $S \rightarrow T$, preventing the visual-text alignment from being trapped in specific narrative patterns.

3.3 Counterfactual Intervention via Hard Negatives

Motivation. The most critical challenge is the Visual Context Confounder ($E \rightarrow A$). To eliminate this confounding effect and strengthen the true

Causal Link ($A \rightarrow I$), we perform Counterfactual Intervention. We achieve this through Generative Hard Negative Mining based on counterfactuals.

Counterfactual Synthesis. We synthesize two types of hard negatives to approximate the intervention $do(A)$:

1) Action-Specific Counterfactuals (Cutting $E \rightarrow A$): We synthesize captions describing a *conflicting* action A' within the *same* visual environment E (*e.g.*, altering "sliding down a water slide" to "jumping on a water slide"). By introducing instances where E is present but A is absent, we explicitly break the spurious correlation (*e.g.*, that "water slide implies sliding"). This penalizes the model for relying on environmental shortcuts and forces it to isolate the true action.

2) Environment-Specific Counterfactuals (Refining $A \rightarrow I$): We preserve the *same* action A but *swap* visual attributes (*e.g.*, clothing color or static background). This intervention breaks the spurious link between the environment and the label, ensuring the established causal link $A \rightarrow I$ is robustly grounded in fine-grained behavioral cues rather than static environmental noise.

By optimizing against these counterfactuals, we force the model to explicitly disentangle causal behavioral cues from environmental noise.

3.4 Optimization Objectives

Our framework intervenes purely at the training data distribution and loss landscape. This design ensures compatibility with existing VLM pipelines and scales linearly with dataset size. Without loss of generality, we adopt a competitive baseline, *i.e.*, X^2-VLM (Zeng et al., 2024) as our backbone, which comprises vision, text, and fusion modules (Figure 3). The training integrates basic multimodal objectives with our causal intervention loss.

Base Objectives. Following X^2-VLM, the backbone is optimized via three standard losses:

1) Image-Text Contrastive (ITC) Loss aligns normalized visual ($\mathbf{v}$) and textual ($\mathbf{t}$) features via symmetric cross-entropy with temperature τ:

$$\mathcal{L}_{itc} = -\frac{1}{2}\mathbb{E}_{(I,T)}\left[\log \frac{e^{\mathbf{v}^T\mathbf{t}/\tau}}{\sum_{T'} e^{\mathbf{v}^T\mathbf{t}'/\tau}} + \log \frac{e^{\mathbf{t}^T\mathbf{v}/\tau}}{\sum_{I'} e^{\mathbf{t}^T\mathbf{v}'/\tau}}\right], \tag{2}$$

2) Image-Text Matching (ITM) Loss predicts the matching probability $\hat{p}$ (with ground-truth y) using the fusion module, enhanced by *in-batch hard negative mining*:

$$\mathcal{L}_{itm} = \mathbb{E}_{(I,T)\sim\mathcal{D}_{hard}}\left[-y\log\hat{p} - (1-y)\log(1-\hat{p})\right], \tag{3}$$

3) Masked Language Modeling (MLM) Loss reconstructs randomly masked text tokens ($\hat{T}$) conditioned on the image I:

$$\mathcal{L}_{mlm} = -\mathbb{E}_{(I,\hat{T})}\left[\sum_{i\in\text{mask}} \log P(t_i|\hat{T}, I)\right], \tag{4}$$

Causal Intervention Loss (Ours). While ITM uses in-batch negatives, random batch samples rarely provide the fine-grained *counterfactual* conflicts needed to correct causal biases. To bridge this gap, we propose a Generative Hard Negative Mining Loss ($\mathcal{L}_{hard}$). For an image I, we construct a triplet comprising a positive text T_{pos} and two synthesized counterfactuals: the action-specific (T_{act}) and environment-specific (T_{env}) negatives (Figure 3). At each training iteration, T_{pos} is randomly sampled from the augmented set $\mathcal{T}_{pos}$ (Section 3.2) to continuously decouple narrative styles. We optimize the model to discriminate the positive pair from these semantic confounders using a categorical cross-entropy loss:

$$\mathcal{L}_{hard} = -\log\left(\frac{\exp(s(I,T_{pos})/\tau_{hard})}{\exp(s(I,T_{pos})/\tau_{hard}) + \sum_{k\in\{act,env\}}\exp(s(I,T_k)/\tau_{hard})}\right) \tag{5}$$

where τ_{hard} (*e.g.*, 0.05) sharpens the decision boundary. By aggregating scores from both counterfactuals ($k \in \{act, env\}$) in the denominator, this loss explicitly penalizes the model for exploiting background context (T_{act}) or static environment features (T_{env}) as shortcuts when the core action mismatches.

Causal Interpretation of $\mathcal{L}_{hard}$. Our causal intervention loss can be formally grounded in the backdoor adjustment criterion. In the observational SCM (Figure 2(a)), the target causal link $A \rightarrow I$ is confounded by the backdoor path $A \leftarrow E \rightarrow I$. Causal identification requires estimating $P(I \mid do(A)) = \sum_e P(I \mid A, e)P(e)$, which marginalizes over the environment E. Standard contrastive objectives implicitly optimize $P(I \mid A, E)$, retaining environmental shortcuts. In contrast, $\mathcal{L}_{hard}$ (Eq. 5) approximates the backdoor adjustment through structured negative sampling. By explicitly constructing T_{act} (holding E fixed, varying A) and T_{env} (holding A fixed, varying E), the loss creates a localized probability distribution that penalizes spurious correlations. Optimizing

$\mathcal{L}_{hard}$ minimizes the conditional mutual information $\mathcal{I}(f(I); E \mid A)$ while maximizing $\mathcal{I}(f(I); A)$, effectively driving the encoder toward representations that satisfy $f(I) \perp E \mid A$. The temperature parameter τ_{hard} further sharpens the gradient, ensuring that the model learns the invariant causal effect $P(I \mid do(A))$ rather than the observational likelihood $P(I \mid A, E)$. Thus, $\mathcal{L}_{hard}$ serves as a differentiable proxy for causal deconfounding within a contrastive learning objective.

Total Objective. The final training objective is a weighted sum:

$$\mathcal{L}_{total} = \mathcal{L}_{itc} + \mathcal{L}_{itm} + \mathcal{L}_{mlm} + \lambda \mathcal{L}_{hard}. \quad (6)$$

where λ is the weight of $\mathcal{L}_{hard}$.

4 Experiments

We conduct comprehensive experiments on the large-scale PAB benchmark and an unseen (OOD) test set to evaluate the effectiveness of our approach. Specifically, our empirical analysis aims to answer the following 3 research questions:

- **RQ1:** How does our proposed dual generative causal intervention framework perform on the fine-grained PAB retrieval task compared with state-of-the-art baselines?
- **RQ2:** How does each component of our method (*i.e.*, Semantic Intervention via Multi-Caption and Counterfactual Intervention via Hard Negatives) contribute to the retrieval performance?
- **RQ3:** How is the generalization and transferability of our model when facing unseen (OOD) scenarios?

4.1 Experimental Setup

Datasets. We evaluate our framework on two publicly available benchmarks to assess in-domain performance and out-of-distribution (OOD) generalization:

- **PAB Benchmark** (Yang et al., 2025b): Used for in-domain evaluation. We train on the 0.1M subset for efficiency and test on the standard 1,978 real-world image-text pairs.
- **UCC (OOD Test Set)** (Yang et al., 2025b): Comprising 5,320 image-text pairs extracted from UCF-Crime (Sultani et al., 2018). We use UCC strictly for *zero-shot* evaluation to rigorously verify transferability to unseen domains.

Evaluation Metrics. Following standard protocols (Jiang and Ye, 2023; Yang et al., 2025b), we measure text-to-image retrieval performance using Recall at K (R@1, R@5, R@10) and mean Average Precision (mAP).

Implementation Details. We implement our framework in PyTorch with an X^2-VLM (Zeng et al., 2024) backbone on a single NVIDIA A100 (80GB) GPU. Images are resized to 224×224. Models are trained for 100 epochs using the AdamW optimizer (Loshchilov and Hutter, 2017) (weight decay 0.01) with a cosine-decayed learning rate starting at 1×10^{-4}. For causal intervention, each image is expanded with 5 augmented positives and 2 counterfactual hard negatives (action- and environment-specific). Causal loss hyperparameters are set to $\tau_{hard} = 0.05$ and $\lambda = 0.7$.

4.2 Performance Comparison (RQ1)

Tables 1 and 2 summarize our comparisons on the PAB and OOD-UCC benchmarks.

Performance on PAB Benchmark (In-Domain). General VLMs (*e.g.*, CLIP (Radford et al., 2021)) and specialized methods (*e.g.*, IRRA (Jiang and Ye, 2023), RaSa (Bai et al., 2023)) struggle with fine-grained anomalies, yielding suboptimal R@1 (Table 1). While the X^2-VLM (Zeng et al., 2024) baseline is competitive, it remains hindered by environmental confounders.

Conversely, our framework significantly elevates all metrics, achieving 89.73% R@1 (+1.71%) and 94.56% mAP (+1.01%) over X^2-VLM (Zeng et al., 2024). This confirms that our causal intervention helps the model to prioritize subtle behavioral cues (*e.g.*, "falling" vs. "lying") over spurious background correlations, ensuring a robust visual-semantic alignment.

Performance on UCC Benchmark (OOD). To assess generalization under distribution shifts, we conduct a zero-shot evaluation on the unseen UCC dataset (Table 2). While most methods suffer severe performance drops due to overfitting to the source domain, our framework demonstrates superior transferability, achieving 73.74% R@1 to outperform the baseline. This firmly confirms our causal analysis: explicitly mitigating Narrative Style Bias and deconfounding the Visual Context Confounder enables the model to capture invariant causal behavioral cues, ensuring robustness across diverse datasets.

Table 1: Performance comparison on the **PAB benchmark** (In-Domain). The best results are highlighted in **bold**.

Method	R@1	R@5	R@10	mAP
Pure pretraining				
MRA (Yang et al., 2025a)	9.91	23.66	31.45	17.15
RaSa (Bai et al., 2023)	21.74	27.30	27.96	24.35
WoRA (Sun et al., 2025)	22.25	45.91	53.54	33.39
APTM (Yang et al., 2023)	22.90	45.80	52.38	33.56
CAMeL (Yu et al., 2025)	24.47	50.00	58.75	36.75
IRRA (Jiang and Ye, 2023)	30.59	59.61	68.91	44.41
CLIP (Radford et al., 2021)	47.57	81.55	89.03	62.73
Pretrain-then-Finetune				
MRA (Yang et al., 2025a)	70.53	94.69	97.47	81.59
APTM (Yang et al., 2023)	72.14	95.30	97.17	82.78
CAMeL (Yu et al., 2025)	74.30	96.79	98.84	84.20
WoRA (Sun et al., 2025)	74.47	96.82	98.48	84.60
IRRA (Jiang and Ye, 2023)	76.39	97.62	99.14	86.33
CLIP (Radford et al., 2021)	77.60	98.84	99.75	87.35
RaSa (Bai et al., 2023)	80.79	98.89	99.65	89.20
CMP (Yang et al., 2025b)	82.62	98.58	99.34	90.14
CMP + **Ours**	86.55 (+3.93)	99.09 (+0.51)	99.34 (+0)	92.48 (+2.34)
X^2-VLM (Zeng et al., 2024)	88.02	99.60	99.80	93.55
X^2-VLM + **Ours**	**89.73** (+1.71)	**99.69** (+0.09)	**99.79** (-0.01)	**94.56** (+1.01)

Table 2: Generalization performance on the unseen **UCC (UCF-Crime) test set** (OOD). The best results are highlighted in **bold**.

Method	R@1	R@5	R@10	mAP
APTM (Yang et al., 2023)	27.86	40.41	46.77	22.61
IRRA (Jiang and Ye, 2023)	40.28	57.24	65.98	33.53
CLIP (Radford et al., 2021)	51.60	68.31	76.43	43.05
RaSa (Bai et al., 2023)	54.12	70.32	75.96	39.71
CMP (Yang et al., 2025b)	52.01	67.53	73.32	39.65
CMP + **Ours**	54.62 (+2.61)	67.55 (+0.02)	73.74 (+0.42)	40.63 (+0.98)
X^2-VLM (Zeng et al., 2024)	70.51	82.99	87.41	54.17
X^2-VLM + **Ours**	**73.74** (+3.23)	**85.62** (+2.63)	**89.47** (+2.06)	**56.63** (+2.46)

Table 3: Ablation study of different components on the PAB dataset. **M-Cap** denotes Multi-Caption Augmentation; **EDA** denotes Easy Data Augmentation; **Env Neg** is Environment-specific Hard Negatives; **Act Neg** denotes Action-specific Hard Negatives.

	Components				Metrics (%)			
	M-Cap	EDA	Env Neg	Act Neg	R@1	R@5	R@10	mAP
1	-	-	-	-	88.02	99.60	99.80	93.55
2	✓	-	-	-	88.98	**99.75**	99.85	94.14
3	✓	✓	-	-	88.82	**99.75**	**99.90**	94.10
4	✓	-	✓	-	88.93	99.70	**99.90**	94.16
5	✓	-	-	✓	88.32	99.65	99.85	93.76
6	✓	-	✓	✓	**89.73**	99.69	99.79	**94.56**

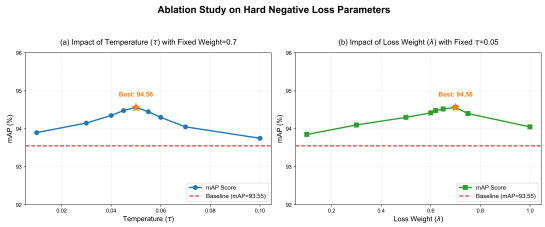

Figure 4: **Hyperparameter sensitivity on PAB.** mAP (%) under varying **(a)** temperature τ_{hard} and **(b)** causal loss weight λ. Peak performance (94.56%, starred) at $\tau_{hard} = 0.05$ and $\lambda = 0.7$ demonstrates the necessity of a sharp decision boundary and a balanced causal constraint.

4.3 Ablation Studies (RQ2)

To answer RQ2, we conduct a step-by-step ablation study on the PAB dataset (Table 3) to verify the contributions of Semantic Intervention (Multi-Caption Augmentation) and Counterfactual Intervention (Hard Negatives).

Effectiveness of Semantic Intervention. Expanding the single caption to 5 augmented captions (M-Cap, Row 2) boosts R@1 from 88.02% to 88.98% over the baseline. This confirms that marginalizing over diverse linguistic styles effectively mitigates *narrative style bias*, enabling the model to focus on invariant semantic content.

Ineffectiveness of Traditional Data Augmentation. Integrating Easy Data Augmentation (EDA, Row 3) with M-Cap drops R@1 to 88.82%. Traditional word-level random modifications (*e.g.*, random deletion) often destroy the delicate semantic structure of anomalous actions. This highlights the necessity of our Generative Causal Intervention (Row 6), which carefully manipulates counterfactuals while preserving core logical structures.

Synergy of Counterfactual Interventions. Interestingly, partial intervention (adding only Env Neg or Act Neg in Rows 4-5) slightly degrades R@1 compared to M-Cap alone (88.98%). We attribute this to *causal entanglement*: since visual context and action semantics are inextricably linked, penalizing only one factor causes the model to over-correct or struggle with the remaining unintervened confounders. Thus, a single-sided constraint is insufficient to establish a robust causal link $A \rightarrow I$.

Conversely, applying both interventions simultaneously (Row 6) yields peak performance (89.73% R@1, 94.56% mAP). This strong synergy confirms that decoupling both potential confounders is vital for isolating true behavioral cues, making every framework component indispensable.

4.4 Sensitivity Analysis (RQ3)

To address RQ3, we investigate our framework's robustness across varying hyperparameters. Notably, all configurations consistently outperform the baseline (93.55% mAP), demonstrating the inherent stability of our causal intervention.

Impact of Hard Negative Temperature (τ_{hard}). Fixing $\lambda = 0.7$, we vary τ_{hard} (Figure 4a). Performance peaks at 94.56% mAP when $\tau_{hard} = 0.05$. Higher temperatures (> 0.05) drop accuracy below

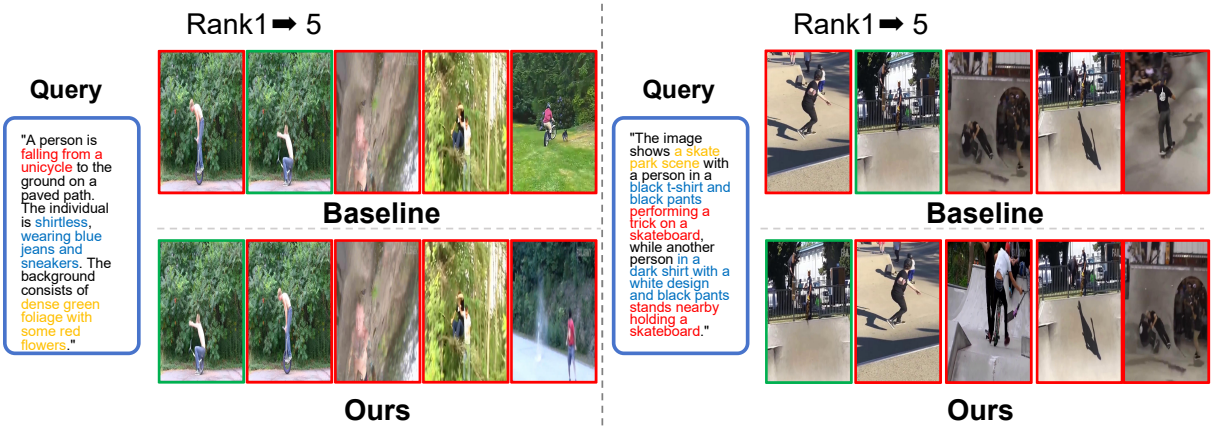

Figure 5: **Qualitative top-5 retrieval results on PAB. (Left)** The baseline is misled by the background context, retrieving normal riding actions (red boxes), whereas Ours correctly identifies the "falling" anomaly (green boxes). **(Right)** The baseline retrieves generic skating within the correct scene but misses the specific "trick". This confirms our framework effectively eliminates visual context confounders to capture fine-grained actions.

94.0%, as overly smooth distributions fail to penalize high-similarity hard negatives. Conversely, an extremely low temperature (0.01) yields a suboptimal $\sim 93.9\%$, indicating overly aggressive gradients. Thus, $\tau_{hard} = 0.05$ strikes the optimal balance, providing gradients sharp enough to effectively distinguish subtle anomalies.

Impact of Causal Loss Weight (λ). Fixing $\tau_{hard} = 0.05$, we sweep λ (Figure 4b). The model achieves its global maximum at $\lambda = 0.7$ (94.56% mAP). A small weight (0.1) results in lower performance ($\sim 93.8\%$), as the causal signal is overwhelmed by base losses. Conversely, pushing λ to 1.0 drops performance to $\sim 94.0\%$. This indicates *over-regularization*, where excessive focus on hard negatives distorts the general visual-semantic alignment. Therefore, $\lambda = 0.7$ establishes the optimal equilibrium.

4.5 Qualitative Analysis

Figure 5 illustrates the top-5 retrieval results for two challenging queries, providing an intuitive understanding of how our framework corrects visual-semantic misalignments.

Case 1: Deconfounding Action from Background. For the query "*falling from a unicycle*" (left), the baseline (Zeng et al., 2024) overfits to spurious environmental correlations ($E \rightarrow I$). It incorrectly retrieves normal riding actions simply because they share the dominant green background, demonstrating a clear vulnerability to visual shortcuts. Conversely, our model correctly prioritizes subtle motion cues to retrieve true "falling" anomalies despite the distracting foliage, further validating the efficacy of Action-Specific Counterfactuals.

Case 2: Distinguishing Fine-grained Attributes. For the query "*performing a trick on a skateboard*" (right), the baseline recognizes the overarching skate park scene but misses the fine-grained action ($E \rightarrow A$), retrieving people merely standing or riding normally without complex dynamic characteristics. In contrast, our method accurately retrieves the specific trick (Ranks 1-2) while simultaneously matching the correct clothing attributes, effectively demonstrating our framework's capacity to cleanly disentangle subtle behavioral semantics from coarse environmental contexts.

5 Conclusion

In this paper, we study the generalization failure in fine-grained pedestrian anomaly behavior retrieval from a causal perspective. We identify that existing methods are hindered by Narrative Style Bias and Visual Context Confounders, leading to models that rely on spurious correlations rather than true behavioral semantics. To mitigate these issues, we propose a plug-and-play dual generative causal intervention framework. Specifically, we implement Semantic Intervention via Multi-Caption Augmentation to enrich the linguistic embedding space, and Counterfactual Intervention via LMM-synthesized Action-Specific and Environment-Specific hard negatives to explicitly decouple causal behavioral cues from environmental noise. Extensive experiments on PAB and the unseen OOD dataset (UCC) validate the effectiveness, superiority, and robust transferability of our proposed paradigm in distinguishing subtle anomalous actions.

6 Limitations

A primary limitation of our framework lies in its reliance on LMMs to synthesize interventional data. Inherent biases or hallucinations within these foundation models can propagate into the generated captions. For instance, in complex pedestrian scenes, the LMM might misinterpret visual details, inadvertently producing noisy counterfactuals (e.g., false negatives) that compromise the causal training objective. Importantly, this limitation stems from the visual grounding bottleneck of current off-the-shelf LMMs, rather than a theoretical flaw in our causal decoupling mechanism. Additionally, synthesizing dense multi-caption sets introduces extra computational overhead during offline data preparation. Future work will explore automated noise-filtering mechanisms and leverage evolving LMM architectures to naturally mitigate these biases and further enhance performance.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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. *arXiv preprint arXiv:2305.13653*.

Chen Chen, Bowen Zhang, Liangliang Cao, Jiguang Shen, Tom Gunter, Albin Jose, Alexander Toshev, Yantao Zheng, Jonathon Shlens, Ruoming Pang, and Yinfei Yang. 2023a. STAIR: Learning sparse text and image representation in grounded tokens. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15079–15094, Singapore. Association for Computational Linguistics.

Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. 2023b. mCLIP: Multilingual CLIP via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13028–13043, Toronto, Canada. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Neil Chowdhury, Franklin Wang, Sumedh Shenoy, Douwe Kiela, Sarah Schwettmann, and Tristan Thrush. 2024. Nearest neighbor normalization improves multimodal retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22571–22582, Miami, Florida, USA. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Zhongjie Duan, Chengyu Wang, Cen Chen, Wenmeng Zhou, Jun Huang, and Weining Qian. 2023. Match4match: Enhancing text-video retrieval by maximum flow with minimum cost. In *Proceedings of the ACM Web Conference 2023*, pages 3257–3267.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in neural information processing systems*, 33:6616–6628.

Zixin Guo, Tzu-Jui Julius Wang, Selen Pehlivan, Abduljalil Radman, and Jorma Laaksonen. 2023. Pitl: cross-modal retrieval with weakly-supervised vision-language pre-training via prompting. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2261–2265.

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

Ding Jiang and Mang Ye. 2023. 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.

Zaid Khan, BG Vijay Kumar, Xiang Yu, Samuel Schulter, Manmohan Chandraker, and Yun Fu. 2022. Single-stream multi-level alignment for vision-language pretraining. In *European Conference on Computer Vision*, pages 735–751. Springer.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11336–11344.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

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

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Shuaiyang Li, Dan Guo, Kang Liu, Richang Hong, and Feng Xue. 2023b. Multimodal counterfactual learning network for multimedia-based recommendation. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 1539–1548.

Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person search with natural language description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1970–1979.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, and 1 others. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European conference on computer vision*, pages 121–137. Springer.

Haoliang Liu, Tan Yu, and Ping Li. 2021. Inflate and shrink:enriching and reducing interactions for fast text-image retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9796–9809, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30.

Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2023. End-to-end knowledge retrieval with multi-modal queries. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8573–8589, Toronto, Canada. Association for Computational Linguistics.

Mingyuan Mao, Renrui Zhang, Honghui Zheng, Teli Ma, Yan Peng, Errui Ding, Baochang Zhang, Shumin Han, and 1 others. 2021. Dual-stream network for visual recognition. *Advances in Neural Information Processing Systems*, 34:25346–25358.

Leland Gerson Neuberg. 2003. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685.

Kai Niu, Yanyi Liu, Yuzhou Long, Yan Huang, Liang Wang, and Yanning Zhang. 2024. An overview of text-based person search: Recent advances and future directions. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(9):7803–7819.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Judea Pearl. 2010. Causal inference. *Causality: objectives and assessment*, pages 39–58.

James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. 2008. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Shivaen Ramshetty, Gaurav Verma, and Srijan Kumar. 2023. Cross-modal attribute insertions for assessing the robustness of vision-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15974–15990, Toronto, Canada. Association for Computational Linguistics.

Alireza Salemi, Juan Altmayer Pizzorno, and Hamed Zamani. 2023. A symmetric dual encoding dense retrieval framework for knowledge-intensive visual question answering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 110–120.

Taoyu Su, Jiawei Sheng, Duohe Ma, Xiaodong Li, Juwei Yue, Mengxiao Song, Yingkai Tang, and Tingwen Liu. 2025. Mitigating modality bias in multi-modal entity alignment from a causal perspective. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1186–1196.

Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 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 *Proceedings of the ACM on Web Conference 2025*, pages 2341–2351.

Yaxiong Wang, Hao Yang, Xiuxiu Bai, Xueming Qian, Lin Ma, Jing Lu, Biao Li, and Xin Fan. 2020. Pfan++: Bi-directional image-text retrieval with position focused attention network. *IEEE Transactions on Multimedia*, 23:3362–3376.

Yihao Wang, Meng Yang, and Rui Cao. 2024. Fine-grained semantic alignment with transferred person-sam for text-based person retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5432–5441.

Dongqing Wu, Huihui Li, Cang Gu, Lei Guo, and Hang Liu. 2024. Dual stream relation learning network for image-text retrieval. *IEEE Transactions on Multimedia*.

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. 2025b. Beyond walking: A large-scale image-text benchmark for text-based person anomaly search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11720–11730.

Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. 2023. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM international conference on multimedia*, pages 4492–4501.

Linli Yao, Weijing Chen, and Qin Jin. 2023. Capenrich: Enriching caption semantics for web images via cross-modal pre-trained knowledge. In *Proceedings of the ACM Web Conference 2023*, pages 2392–2401.

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*.

Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2024. $X^2$2-vlm: All-in-one pre-trained model for vision-language tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3156–3168.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024a. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644.

Peng-Fei Zhang, Zi Huang, and Guangdong Bai. 2024b. Universal adversarial perturbations for vision-language pre-trained models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 862–871.

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

A LMM Prompting Details for Generative Causal Intervention

To ensure reproducibility, we detail the prompt templates used for our LMM pipeline. Throughout our experiments, counterfactual hard negative syntheses are powered by Qwen3-VL-8B. We employ a multi-turn dialogue strategy to strictly constrain the LMM's output, ensuring it decouples causal links without introducing hallucinated objects. For brevity, we present the core instructional snippets below.

A.1 Prompt for Action-Specific Counterfactuals

To effectively break the causal link between the visual context confounder and the anomaly action ($E \rightarrow A$), the generated negative action must be a completely physically impossible or semantically opposite action given the context, rather than a mere synonym.

Design Rationale: Absurd vs. Plausible Counterfactuals. An intuitive approach to generating action negatives would be selecting plausible but incorrect fine-grained actions (*e.g.*, generating "tripping" instead of "falling"). However, during early explorations, we found that this strategy severely harms the contrastive optimization. In fine-grained anomaly datasets, the visual boundary between "tripping" and "falling" is highly ambiguous in static frames. Using such plausible actions often results in *False Negatives*—where the generated negative text actually matches the image visually. Forcing the model to push away these false negatives confuses the alignment space. Therefore, we explicitly instruct the LMM to generate actions that are *physically impossible, absurd, or mutually exclusive* (*e.g.*, "levitating three feet above the ground" instead of "falling"). This ensures a definitive *True Negative* boundary, providing a clean gradient for the model to learn that the precise physical posture of "falling" is completely distinct from the counterfactual action.

[System]: EXPERT INSTRUCTION FOR CREATING NEGATIVE SAMPLES. Your task is to generate captions with actions that are: 1. PHYSICALLY IMPOSSIBLE or EXTREMELY UNLIKELY in this context. 2. Semantically OPPOSITE to the original action. 3. Obviously wrong to a human observer.
CRITICAL RULES: NEVER use synonyms (*e.g.*, 'holding'→'carrying'). ALWAYS choose actions requiring a COMPLETELY DIFFERENT PHYSICAL POSTURE. [...]
[User]: <*Image Input*> Original caption: {original_caption}
[Assistant]: I understand. I will replace the main action to create a COMPLETELY DIFFERENT SCENE without using synonyms, while keeping appearance intact. [...]
[User]: FINAL INSTRUCTIONS:
1. Identify the MAIN ACTION and OBJECT(S).
2. Replace ONLY the main action with something PHYSICALLY IMPOSSIBLE or ABSURD, requiring a DIFFERENT body posture.
3. Keep ALL appearance details EXACTLY the same.
OUTPUT RULES: Output ONLY the new caption with the same grammatical structure. Do NOT add or remove objects.

A.2 Prompt for Environment-Specific Counterfactuals

To refine the causal link from action to the image ($A \rightarrow I$) and prevent the model from using clothing or colors as shortcuts, we strictly lock the action verbs while perturbing the environment details.

[System]: You are an expert at modifying captions while preserving specific elements. Your task is to modify ONLY the appearance details in the given caption while keeping all action verbs exactly the same. [...]
[User]: <*Image Input*> Original caption: {original_caption}
[Assistant]: I understand. I will modify only the appearance details while keeping all action verbs intact.
[User]: Create a new caption where the APPEARANCE DETAILS (clothing colors, object colors, physical attributes) are changed to something plausible but incorrect.
CRITICAL RULE: Keep the main ACTION/VERB EXACTLY the same as the original caption.
OUTPUT RULES: Output ONLY the new caption, no explanations.

B Qualitative Examples of Generated Counterfactuals

To provide an intuitive understanding of our Generative Causal Intervention strategy, we present qualitative examples from the Pedestrian Anomaly Behavior (PAB) dataset augmented by our LMM pipeline in Table 4.

For each image, the baseline dataset provides a single *Original Caption*. Through our Multi-Caption Augmentation, we generate diverse positive descriptions. More importantly, guided by the causal prompts detailed in Appendix A, we synthesize two types of counterfactual hard negatives:

- **Action-Specific Negatives (Act Neg):** The model replaces the core action with a physically impossible, absurd, or semantically opposite action, while keeping all appearance and background details exactly the same. For example, replacing "holding a spider web" with "violently smashing the spider web with a hammer".

- **Appearance-Specific Negatives (Env Neg):** The model swaps clothing colors, object attributes, and backgrounds to plausible but incorrect details, while strictly preserving the main action verbs. For example, changing a "black punching bag" to a "pink punching bag with glittery purple lettering".

These counterfactual examples compel the retrieval model to explicitly deconfound behavioral cues from the visual context, heavily penalizing models that rely on superficial keyword matching.

C Data Generation Pipeline: Efficiency, Quality Control, and Statistics

C.1 Computational Overhead

A common concern regarding Generative Causal Intervention is the computational bottleneck of utilizing Large Multimodal Models (LMMs). In our framework, the counterfactual synthesis is treated entirely as an offline pre-processing step. Generating the augmented dataset (5 positive multi-captions and 2 counterfactual hard negatives per image) for the entire 0.1M PAB training subset required approximately 72 hours (3 days) on a single NVIDIA RTX 4090 (24GB) GPU using the Qwen3-VL-8B model. This indicates that our data-centric intervention is highly accessible and does not add any computational overhead to the actual training or inference phases of the retrieval model.

C.2 Quality Control and Filtering Strategy

While LMMs are powerful, their raw outputs often contain conversational artifacts (*e.g.*, "Here is the modified caption:", or internal thinking processes). To ensure the purity of the semantic space, we deployed a rigorous post-processing filtering strategy:

1. **Rule-based Parsing:** We utilize regular expressions to strip conversational prefixes and suffixes, strictly extracting the core descriptive strings. Captions containing refusal phrases (*e.g.*, "I cannot generate...") were filtered and regenerated.

2. **Human Evaluation:** We randomly sampled 100 generated positive-negative triplets for manual inspection. The human evaluation confirmed high quality: the counterfactuals logically manipulated the target attributes (action or appearance) without breaking the overall grammatical structure. Even the deliberately "absurd" actions (*e.g.*, levitating) integrated seamlessly into the syntax of the original caption, providing an excellent challenging negative for contrastive learning without appearing as garbled text.

C.3 Text Token Length Analysis

To further verify the consistency of the filtered captions, we analyze the text token lengths. As illustrated in Figure 6, the word count distribution exhibits a well-balanced normal distribution. The average length is 58.78 words, and the 99th percentile reaches 100 words. This distribution ensures **High Information Density** (covering complex anomaly details) and **Length Consistency** (preventing the model from learning a spurious "length bias"). Furthermore, a 99th percentile of 100 words guarantees that the vast majority of our captions fit within the maximum token limit of standard Vision-Language Models (*e.g.*, X^2-VLM), avoiding severe semantic truncation.

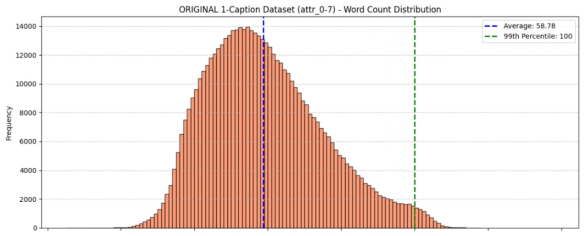

Figure 6: **Word count distribution of the generated captions.** The lengths exhibit a normal distribution with an average of 58.78 words, ensuring consistent information density without severe truncation issues during tokenization.

D Extended Analysis on the Synergy of Counterfactual Interventions

In Section 4.3 (Table 3), we reported a counter-intuitive phenomenon: applying *partial intervention* (only Env Neg or only Act Neg) leads to a slight performance drop compared to using Multi-Captions alone. We provide a deeper causal explanation for this behavior here.

Table 4: Qualitative examples of the generated counterfactual hard negatives. Differences from the original caption are highlighted in **bold red** for Action Negatives and **bold blue** for Environment/Appearance Negatives.

Scene Image	Data Type	Caption Text
 Example 1 *Lawn* *(Spider Web)*	Original Pos.	Two girls are in a grassy area. One girl is kneeling on the grass, wearing a light gray long-sleeve shirt, dark pants, and light-colored shoes. She is **holding a large, intricate spider web**. The other girl is standing, wearing a light gray short-sleeve shirt, black pants, and brown shoes. She is holding a black object, possibly a tool or a stick. The background includes a wooden structure and a tree.
	Action Neg.	Two girls are in a grassy area. One girl is kneeling on the grass, wearing a light gray long-sleeve shirt, dark pants, and light-colored shoes. She is **violently smashing the spider web with a hammer**. The other girl is standing, wearing a light gray short-sleeve shirt, black pants, and brown shoes. She is holding a black object, possibly a tool or a stick. The background includes a wooden structure and a tree.
	Env Neg.	Two girls are in a grassy area. One girl is kneeling on the grass, wearing a **bright neon yellow long-sleeve shirt, lime green pants, and neon pink sneakers**. She is holding a large, intricate spider web. The other girl is standing, wearing a **royal blue short-sleeve shirt, glittery silver pants, and mismatched purple and teal boots**. She is holding a **rainbow-striped object**... The background includes a **brick structure and a palm tree**.
 Example 2 *Gym* *(Boxing)*	Original Pos.	A man is **practicing boxing** in a gym. He is wearing a black t-shirt, black shorts, and black boxing gloves. He is **standing in front of** a large black punching bag with gold lettering. The gym has large windows showing a parking lot with several cars.
	Action Neg.	A man is practicing boxing in a gym. He is wearing a black t-shirt, black shorts, and black boxing gloves. He is standing in front of a large black punching bag with gold lettering. The gym has large windows showing a parking lot with several cars. **He is simultaneously levitating three feet above the ground while screaming in terror at the punching bag, which is now glowing with neon lights and emitting smoke**.
	Env Neg.	A man is practicing boxing in a gym. He is wearing a **red t-shirt, neon green shorts, and silver boxing gloves**. He is standing in front of a **large pink punching bag with glittery purple lettering**. The gym has large windows showing a parking lot with several cars.

In the observational PAB dataset, the visual environment (E) and the anomaly action (A) are strongly entangled.

- When we introduce only Environment Negatives (Env Neg), we force the model to become overly invariant to the background. However, because the model still struggles to clearly define the fine-grained action boundaries (due to the lack of Act Neg), stripping away the environmental context removes a "shortcut" the model was relying on. Without a strong signal to learn the action itself, the model becomes confused, leading to optimization instability.

- Conversely, when we introduce only Action Negatives (Act Neg), the model is heavily penalized for confusing actions, but the background remains a reliable constant. The model adopts a lazy strategy: it memorizes specific *background-action pairings* rather than genuinely understanding the action. This over-regularization distorts the embedding space.

Applying both interventions simultaneously breaks this "seesaw effect." It forces the model into a saddle point where it can neither rely solely on the background nor ignore it entirely. The joint constraint compels the embedding space to orthogonally decouple the action features from the appearance features, which is why the full framework achieves a significant performance rebound.

E Comparison with Standard Hard Negative Mining

A natural question arises as to why we design a specific Generative Causal Intervention Loss ($\mathcal{L}_{hard}$) rather than employing standard hard negative mining techniques, such as the Triplet Margin Loss widely used in retrieval tasks.

The standard Triplet Loss focuses on pushing a mined negative T_{neg}^{mined} away from the anchor I by a fixed margin α:

$$\mathcal{L}_{triplet} = \max(0, s(I, T_{neg}^{mined}) - s(I, T_{pos}) + \alpha) \tag{7}$$

In contrast, our proposed objective is formulated as a localized InfoNCE loss over a specifically constructed counterfactual subset:

$$\mathcal{L}_{hard} = -\log\left(\frac{\exp(s(I,T_{pos})/\tau_{hard})}{\exp(s(I,T_{pos})/\tau_{hard}) + \sum_{k\in\{act,env\}} \exp(s(I,T_k)/\tau_{hard})}\right) \tag{8}$$

Our $\mathcal{L}_{hard}$ exhibits three fundamental advantages over standard methods for fine-grained anomaly retrieval:

1. **Generative vs. In-batch Mined:** Standard methods mine hard negatives from the mini-batch. In fine-grained tasks, batch negatives are either too easy (completely different scenes) or inadvertently represent False Negatives (a different image that actually matches the text). Our negatives are *generatively controlled* to differ by exactly one causal variable, ensuring they are definitive True Negatives.

2. **Joint Optimization over Confounders:** Triplet loss treats negatives independently and stops optimizing once the margin α is met. Our $\mathcal{L}_{hard}$ creates a joint probability distribution across the specific confounders (action and environment). This enforces a continuous gradient that simultaneously balances the push against *both* confounders, preventing the "seesaw effect" discussed in Appendix D.

3. **Sharpened Boundary via Temperature:** The inclusion of the low temperature parameter τ_{hard} (0.05) aggressively sharpens the decision boundary. Unlike a static margin α, τ_{hard} forces the model to penalize even slight similarities with the counterfactuals, which is critical for distinguishing highly similar anomalous actions.

F Impact of Semantic Expansion Scale (K)

Due to space constraints in the main text, we provide a more detailed analysis regarding the optimal density for *Semantic Intervention* here. We evaluated our framework with different numbers of augmented positive captions per image, $K \in \{1, 3, 5, 7\}$, where $K = 1$ represents the baseline without expansion. The results are illustrated in Figure 7.

Performance Trend. We observe a consistent and monotonic improvement in R@1 as K increases from 1 to 5. Specifically, the performance jumps from 87.92% ($K = 1$) to a peak of **89.73%** at $K = 5$. However, further increasing the scale to $K = 7$ results in a noticeable performance drop to 88.27%.

Analysis of Improvement ($1 \rightarrow 5$). The significant gain of +1.81% from $K = 1$ to $K = 5$

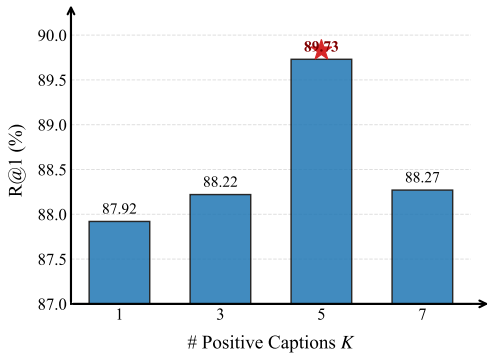

Figure 7: **Impact of Semantic Expansion Scale K.** The retrieval performance (R@1) peaks at $K = 5$, balancing narrative diversity and semantic accuracy.

strongly validates our hypothesis regarding *Narrative Style Bias*. Since original observational annotations often follow a rigid template or human annotators' habitual phrasing, a single caption is insufficient to cover the diverse linguistic structures a user might employ to describe an anomaly in real-world queries. By expanding each image into 5 distinct captions with varying sentence structures and synonyms, the model is forced to marginalize over varying linguistic styles. Consequently, it learns to map diverse texts to the invariant *causal action semantics* rather than overfitting to specific phrasing patterns.

Analysis of Degradation ($5 \rightarrow 7$)**.** The performance drop at $K = 7$ indicates a *semantic saturation point*. We attribute this degradation to two primary factors:

1. *Information Redundancy*: Beyond 5 captions, the generated descriptions for the exact same visual content inevitably become semantically repetitive. The lack of new perspective limits the marginal benefit of adding more positive samples.

2. *LMM Hallucinations*: When the LMM is forced to generate a high volume of distinct captions ($K = 7$) for the same visual input, it may over-extrapolate and introduce subtle hallucinations or noisy details that deviate from the true visual evidence. This noise can disrupt the delicate visual-semantic alignment process during contrastive learning.

Thus, $K = 5$ represents the optimal equilibrium between linguistic diversity and semantic precision, maximizing the deconfounding effect without introducing detrimental noise.